Nonasymptotic quasi-optimality of AIC and the slope heuristics in maximum likelihood estimation of density using histogram models

Adrien Saumard

Nonasymptotic quasi-optimality of AIC and the slope heuristics in maximum likelihood estimation of density using histogram models

A. Saumard
University Rennes 1, IRMAR
adrien.saumard@univ-rennes1.fr

August 29, 2010

Abstract

We consider nonparametric maximum likelihood estimation of density using linear histogram models. More precisely, we investigate optimality of model selection procedures via penalization, when the number of models is polynomial in the number of data. It turns out that the Slope Heuristics first formulated by Birgé and Massart [10] is satisfied under rather mild conditions on the density to be estimated and the structure of the considered partitions. This suggests a new look at AIC penalty and more precisely, we show that the minimal penalty in the sense of Birgé and Massart is equivalent to half AIC penalty. Thus, as soon as the chosen penalty is larger than half AIC, the model selection procedure satisfies an oracle inequality. On contrary, if the penalty is less than the minimal one, then the procedure totally misbehaves. Moreover, if the penalty is equal to AIC penalty - and the number of data is large enough -, then the model selection procedure is nearly optimal in the sense that it satisfies a nonasymptotic, trajectorial oracle inequality with constant almost one, tending to one when the number of data goes to infinity. Finally, it is, to our knowledge, the first time that the Slope Heuristics is theoritically validated in a non-quadratic setting.

Keywords: Maximum likelihood, density estimation, AIC, Optimal model selection, Slope heuristics, Penalty calibration.

1 Introduction

This paper is devoted to the study of some penalized maximum likelihood model selection procedures for the estimation of density on histograms. There is a huge amount of literature on the problem of model selection by penalized maximum likelihood criteria, even in the more restrictive question of selecting an histogram, that goes back to Akaike's pioneer work. In the early seventies, Akaike [1] proposed to select a model by penalizing the empirical likelihood of maximum likelihood estimators by the number of parameters in each model. The analysis of Akaike [1] on the model selection procedure defined by the so-called Akaike's Information Criterion (AIC), is fundamentally asymptotic in the sense that the author considers a given finite collection of models with the number of data going to infinity. This asymptotic setting is irrelevant in many situations and thus many efforts have been made to develop nonasymptotic analysis of model selection procedures, letting the dimension of the models and the cardinality of the collection of models depend on the number of data. As pointed out by Boucheron and Massart [11], it is nevertheless worth mentioning that early works of Akaike [2] and Mallows [22] in model selection relied, although in a disguised form, on the Wilks' phenomenon (Wilks [28]) that asserts that in smooth parametric density estimation the difference between the maximum likelihood and the likelihood of the sampling distribution converges towards a chi-square distribution where the number of degrees of freedom coincides with the model dimension. This phenomenon has been generalized by Boucheron and Massart [11] in a nonasymptotic way, considering the empirical excess risk in a M-estimation with bounded contrast setting, and is actually one the main results supporting the conjecture that the slope heuristics introduced by Birgé and Massart [10] hold in some general framework, see Arlot and Massart [6]. Let us now describe some works related to the selection of maximum likelihood estimators.

Barron and Sheu [9] give some risks bounds on maximum likelihood estimation considering sequences of regular exponential families made of polynomials, splines and trigonometric series. They achieve an accurate

trade-off between the bias term and the variance term considering that log-density functions have square integrable derivatives. Considering general models, Barron, Birgé and Massart [8] give strategies of penalization in a nonasymptotic framework and derive oracle inequalities for the Hellinger risk. In particular, the considered penalty terms take into account the complexity of the collection of models, but as a prize to pay for generality, they involve absolute constants that may be unrealistic.

Particularizing the structure of the models to histograms, Castellan [14] proposes a modified Akaike's criterion that also takes into account the complexity of the collection of models, and that lead to significant changes compared to AIC criterion in the case of large collections of irregular partitions. She derives nonasymptotic oracle inequalities for the Hellinger and Kullback-Leibler risks of the selected model, with leading constants in front of the oracle only depending on the multiplicative constant in the penalty term and being optimized for a penalty term corresponding to AIC in the case of regular histograms. But, despite the fact that she gives optimal controls from above and from below for the mean of the Hellinger and Kullback-Leibler risks on a fixed model (see Proposition 2.4 and 2.6 in [14]), the derived oracle inequalities are not sufficiently sharp to recover the asymptotic optimality of AIC in the case of regular histograms, as the leading constants are bounded away from one even if the number of data is going to infinity. Castellan [14] also give a lower bound for the penalty term that corresponds to half AIC penalty, when the unknown density is uniform on the unit interval and the partitions are regular. This result seems to indicate that the slope heuristics exhibited by Birgé and Massart [10] is satisfied in the context of maximum likelihood estimation of density, at least when the considered models are regular histograms. Castellan [15] has also been able to generalize her study to exponential models where the logarithm of functions are piecewise polynomials. By distinguishing between regular and irregular partitions defining the models, she gives significant bounds in Hellinger risk for procedures of model selection based on a modified Akaike's criterion. We also refer to the introduction of Castellan [14] for a state of the art on the problem of selecting histograms, and in particular the related question of optimal cell width in the case of regular histograms.

We show in this paper that the slope heuristics is valid when the collection of models is of polynomial complexity with respect to the number of data and the considered partitions satisfy some lower regularity assumption. More precisely, we identify the minimal penalty as half AIC penalty. For a penalty function less than the minimal one, we show that the procedure of model selection totally misbehaves in the sense that the Kullback-Leibler excess risk of the selected model is much larger than the oracle one, and the selected dimension is systematically large too. On the contrary, when the penalty function is larger than the minimal one, assuming that the bias of the models are bounded from above and from below by a power of the number of elements in each partition, we show a nonasymptotic pathwise oracle inequality for the Kullback-Leibler excess risk of the selected model. The assumption on the bias of the models holds true when the unknown density is a non constant α-Hölder function. Moreover, if the penalty function is close to two times the minimal one, the leading constant in the oracle inequality is close to one, and is even converging to one when the number of data is going to infinity, meaning that we are close to the optimal penalty. This allows us to show nonasymptotic quasi-optimality of AIC in this context. From a practical point of view, as our results theoretically validate the data-driven calibration of penalty exposed by Arlot and Massart in [6] and as the penalty shape is known in this case and is equal to the dimension of the models, we are able to provide a data-driven model selection procedure that asymptotically behaves like AIC procedure. Moreover, this data-driven procedure should perform better than AIC for small numbers of data. A simulation study about this fact is still in progress.

Our analysis, that significantly differs from Castellan's approach in [14], is based on the concept of contrast's expansion exposed in [24] in the case of least-squares regression. Moreover, on each model, the Kullback-Leibler divergence with respect to the Kullback-Leibler projection of the unknown density is shown to be close to a weighted $L_2(P)$ norm, locally around the Kullback-Leibler projection, where P is the sampling distribution. Our approach then relies on two central facts : under a lower regularity assumption on the partitions, the models are equipped with a localized basis structure, and assuming moreover that the unknown density is uniformly bounded from above, the maximum likelihood estimators are consistent in sup-norm, uniformly over the collection of models, and converge towards their corresponding Kullback-Leibler projections. We notice that this notion of convergence in sup-norm, which is essential in our methodology, is also present in the work of Castellan, slightly disguised in the term $\Omega_m(\varepsilon)$ defined in Section 2.3 of [14].

Finally, histogram models of densities combine two properties : on the one hand they are a particular case of exponential models, and on the other hand they can be viewed as the subset of positive functions in

an affine space. Our approach is based on the second property, whereas Castellan's one relies on the first property, taking advantage of the linear structure of the contrasted functions. We conjecture that the slope phenomenon discovered by Birgé and Massart in a generalized linear Gaussian model setting can be extended in the two directions described above. In each case, one of the main task will be to prove the consistency in sup-norm of the maximum likelihood estimators on the considered models, as further explained in Section 4.

The paper is organized as follows. In Section 2 we describe the statistical framework, the considered models and we investigate in Section 2.3 the "regular" structure of the Kullback-Leibler contrast on histogram models. We state in Section 3 our main results. In Section 4 we give arguments concerning possible developments of the two possible generalizations described above. The proofs are postponed to the end of the paper.

2 Framework and notations

2.1 Maximum Likelihood Estimation

We assume that we have n i.i.d. observations $(\xi_1, ..., \xi_n)$ with common unknown law P on a measurable space $(\mathcal{Z}, \mathcal{T})$ and that ξ is a generic random variable of law P on $(\mathcal{Z}, \mathcal{T})$ and independent of the sample $(\xi_1, ..., \xi_n)$. We also assume that there exists a known probability measure μ on $(\mathcal{Z}, \mathcal{T})$ such that P admits a density s_* with respect to μ :

$$s_* = \frac{dP}{d\mu} \ .$$

Our goal is to estimate the density s_*.
For a measurable suitable integrable function f on $\mathcal{Z}$, we set

$$Pf = P(f) = \mathbb{E}[f(\xi)]$$
$$\mu f = \mu(f) = \int_{\mathcal{Z}} f d\mu$$

and if

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{\xi_i}$$

denote the empirical distribution associated to the data $(\xi_1, ..., \xi_n)$,

$$P_n f = P_n(f) = \frac{1}{n} \sum_{i=1}^{n} f(\xi_i) \ .$$

Moreover, taking the convention $\ln(0) = -\infty$ and defining the positive part as $(x)_+ = x \vee 0$, we set

$$\mathcal{S} = \left\{ s : \mathcal{Z} \longrightarrow \mathbb{R}_+ \ ; \ \int_{\mathcal{Z}} s d\mu = 1 \text{ and } P(\ln s)_+ < +\infty \right\} \ . \tag{1}$$

We assume in the sequel that the unknown density s_* belongs to $\mathcal{S}$. In fact, in order to derive our results, we will assume in Section 3 that s_* is uniformly bounded away from zero and uniformly upper bounded on $\mathcal{Z}$. For now, note that since $P(\ln s_*)_- < +\infty$ and $s_* \in \mathcal{S}$, we have $P|\ln(s_*)| < +\infty$. Moreover, the Kullback-Leibler contrast K is defined on $\mathcal{S}$ to be

$$K : s \in \mathcal{S} \longmapsto (z \in \mathcal{Z} \longmapsto -\ln(s(z)))$$

and thus the risk

$$PK(s) = P(Ks) = PKs = P(\ln s)_- - P(\ln s)_+$$

as well as the excess risk

$$\ell(s_*, s) = P(Ks) - P(Ks_*) = P(Ks - Ks_*)$$

are well defined on $\mathcal{S}$ and can be possibly infinite. Now, for two probability distributions P_s and P_t on $(\mathcal{Z}, \mathcal{T})$ of respective densities s and t with respect to μ, the Kullback-Leibler divergence of P_t with respect to P_s is defined to be

$$\mathcal{K}(P_s, P_t) = \begin{cases} \int_{\mathcal{Z}} \ln\left(\frac{dP_s}{dP_t}\right) dP_t = \int_{\mathcal{Z}} s \ln\left(\frac{s}{t}\right) d\mu & \text{if } P_s \ll P_t \\ +\infty & \text{otherwise.} \end{cases} \tag{2}$$

By misuse of notation we will denote $\mathcal{K}(s,t)$ rather than $\mathcal{K}(P_s, P_t)$ and by Jensen inequality we notice that $\mathcal{K}(s,t)$ is a nonnegative quantity, equal to zero if and only if $s = t \ \ \mu$-$a.s.$ Hence, for any $s \in \mathcal{S}$, the excess risk $\ell(s_*, s)$ satisfies

$$\begin{aligned} \ell(s_*, s) &= P(Ks - Ks_*) \\ &= \int_{\mathcal{Z}} \ln\left(\frac{s_*}{s}\right) s_* d\mu \\ &= \mathcal{K}(s_*, s) \geq 0 \end{aligned} \tag{3}$$

and this nonnegative quantity is equal to zero if and only if $s_* = s \ \ \mu - a.s.$ We thus deduce that the unknown density s_* is uniquely defined by

$$\begin{aligned} s_* &= \arg\min_{s \in \mathcal{S}} \{P(-\ln s)\} \\ &= \arg\min_{s \in \mathcal{S}} \{PK(s)\} \ . \end{aligned} \tag{4}$$

For a subset $\widetilde{M} \subset \mathcal{S}$, we define the maximum likelihood estimator on $\widetilde{M}$, whenever it exists, by

$$\begin{aligned} s_n\left(\widetilde{M}\right) &\in \arg\min_{s \in \widetilde{M}} \{P_n K s\} \\ &= \arg\min_{s \in \widetilde{M}} \left\{ \frac{1}{n} \sum_{i=1}^{n} -\ln(s(\xi_i)) \right\} \ . \end{aligned} \tag{5}$$

Finally, for any $s \in L_2(P)$, we denote by

$$\|s\|_2 = \left(\int_{\mathcal{Z}} s^2 dP \right)^{1/2}$$

its quadratic norm.

2.2 Histogram models

The models $\widetilde{M}$ that we consider here to define the maximum likelihood estimators as in (5) are subsets of linear spaces M made of histograms. More precisely, for a finite partition Λ_M of cardinality $|\Lambda_M| = D_M$, we set

$$M = \left\{ s = \sum_{I \in \Lambda_M} \beta_I \mathbf{1}_I \ ; \ \beta = (\beta_I)_{I \in \Lambda_M} \in \mathbb{R}^{D_M} \right\}$$

the linear vector space of piecewise constant functions with respect to Λ_M and we assume that any element I of the partition Λ_M is of positive measure with respect to μ :

$$\text{for all } I \in \Lambda_M, \quad \mu(I) > 0 \ . \tag{6}$$

By misuse of language, the space M is also called "model" or "histogram model". The linear dimension of M is equal to D_M. In addition we associate to the model M the subset $\widetilde{M}$ of the functions in M that are densities with respect to μ,

$$\widetilde{M} = \left\{ s \in M \ ; \ s \geq 0 \text{ and } \int_{\mathcal{Z}} s d\mu = 1 \right\} \ .$$

As the partition Λ_M is finite, we have $P(\ln s)_+ < +\infty$ for all $s \in \widetilde{M}$ and so $\widetilde{M} \subset \mathcal{S}$. Hence, by (5), we can associate to $\widetilde{M}$ the maximum likelihood estimator $s_n\left(\widetilde{M}\right)$ and in the following we denote it $s_n(M)$ rather than $s_n\left(\widetilde{M}\right)$. We state in the next proposition some well-known properties that are satisfied by histogram models submitted to the procedure of maximum likelihood estimation (see for example Massart [23], Section 7.3).

Proposition 1 *Let*

$$s_M = \sum_{I \in \Lambda_M} \frac{P(I)}{\mu(I)} \mathbf{1}_I \ . \tag{7}$$

Then $s_M \in \widetilde{M}$ and s_M is called the Kullback-Leibler projection of s_* onto $\widetilde{M}$. Moreover, it holds

$$s_M = \arg\min_{s \in \widetilde{M}} P(Ks) \ . \tag{8}$$

The following Pythagorean-like identity for the Kullback-Leibler divergence holds, for every $s \in \widetilde{M}$,

$$\mathcal{K}(s_*, s) = \mathcal{K}(s_*, s_M) + \mathcal{K}(s_M, s) \ . \tag{9}$$

We also have the following formula

$$s_n(M) = \sum_{I \in \Lambda_M} \frac{P_n(I)}{\mu(I)} \mathbf{1}_I \ , \tag{10}$$

and so the maximum likelihood estimator on M is well defined and corresponds to the classical histogram estimator of s_* associated to the partition Λ_M.

Remark 2 *Histogram models are special cases of general exponential families exposed for example in Barron and Sheu [9] (see also Castellan [15] for the case of exponential models of piecewise polynomials). The projection property (9) can be generalized to exponential models (see Lemma 3 of [9] and Csiszár [16]).*

Remark 3 *As by (3) we have*

$$P(Ks_M - Ks_*) = \mathcal{K}(s_*, s_M)$$

and for any $s \in \widetilde{M}$,

$$P(Ks - Ks_*) = \mathcal{K}(s_*, s)$$

we easily deduce from (9) that the excess risk on $\widetilde{M}$ is still a Kullback-Leibler divergence, as we then have for any $s \in \widetilde{M}$,

$$P(Ks - Ks_M) = \mathcal{K}(s_M, s) \ . \tag{11}$$

Moreover it is easy to see using (10) that the maximum likelihood estimator on a histogram model M is also the least-squares estimator.

We shall ask for a particular analytical structure of the considered models in order to derive sharp upper and lower bounds for the excess risk on each model of reasonable dimension. Namely, we require here that the models are fulfilled with a localized basis structure with respect to the $L_2(P)$ norm. As stated in the following lemma, this property is available when the unknown density of data is uniformly bounded away from zero and when the partition Λ_M related to the model M satisfies some lower regularity property with respect to the measure of reference μ.

Lemma 4 *Let $A_{\min}, A_\Lambda > 0$. Let Λ_M be some finite partition of $\mathcal{Z}$ and M be the model of piecewise constant functions on the partition Λ_M. Assume that*

$$\inf_{z \in \mathcal{Z}} s_*(z) \geq A_{\min} > 0 \quad \text{and} \quad D_M \inf_{I \in \Lambda_M} \mu(I) \geq A_\Lambda > 0 \ . \tag{12}$$

Set $r_M = (A_{\min} A_\Lambda)^{-1/2}$ *and define, for all* $I \in \Lambda_M$,

$$\varphi_I = (P(I))^{-1/2} \mathbf{1}_I \ .$$

Then the family $(\varphi_I)_{I \in \Lambda_M}$ *is an orthonormal basis of* $(M, L_2(P))$ *that satisfies, for all* $\beta = (\beta_I)_{I \in \Lambda_M} \in \mathbb{R}^{D_M}$,

$$\left\| \sum_{I \in \Lambda_M} \beta_I \varphi_I \right\|_\infty \leq r_M \sqrt{D_M} \, |\beta|_\infty \tag{13}$$

where $|\beta|_\infty = \max \{ |\beta_I| \ , \ I \in \Lambda_M \}$. *As a consequence,*

$$\sup_{s \in M, \ \|s\|_2 \leq 1} \|s\|_\infty \leq r_M \sqrt{D_M} \ . \tag{14}$$

The proof of Lemma 4 is straightforward and can be found in Section 5.1.

2.3 "Regularity" of the Kullback-Leibler contrast

Our goal is to study the performance of maximum likelihood estimators, that we measure by their excess risk. So we are interested in the random quantity $P(Ks_n(M) - Ks_*)$. Moreover, since we can write

$$P(Ks_n(M) - Ks_*) = P(Ks_n(M) - Ks_M) + P(Ks_M - Ks_*)$$

and since the bias $P(Ks_M - Ks_*)$ is deterministic, we focus on the quantity

$$P(Ks_n(M) - Ks_M) \geq 0 \ ,$$

that we want to bound in probability. We will often call this last quantity the excess risk of the estimator on M or the true excess risk of $s_n(M)$, by opposition to the empirical excess risk for which the expectation is taken over the empirical measure : $P_n(Ks_M - Ks_n(M)) \geq 0$.
We notice that by Proposition 1, the excess risk of the maximum likelihood estimator on M is still a Kullback-Leibler divergence if M is a model of histograms, as we have

$$P(Ks_n(M) - Ks_M) = \mathcal{K}(s_M, s_n(M)) \ .$$

The following lemma provides an expansion of the contrast around s_M on M as the sum of a linear part and a second order part which behaves as a quadratic.

Lemma 5 *Assume that*

$$\inf_{z \in \mathcal{Z}} s_*(z) \geq A_{\min} > 0 \tag{15}$$

and consider $s \in \widetilde{M}$ *such that*

$$\|s - s_M\|_\infty < A_{\min} \ . \tag{16}$$

Then we have $\inf_{z \in \mathcal{Z}} s(z) > 0$ *and it holds for all* $z \in \mathcal{Z}$,

$$(Ks)(z) - (Ks_M)(z) = \psi_{1,M}(z)(s - s_M)(z) + \psi_2\left(\left(\frac{s - s_M}{s_M}\right)(z)\right) \tag{17}$$

with

$$\psi_{1,M}(z) = -\frac{1}{s_M(z)}$$

and, for all $t \in (-1, +\infty)$,

$$\psi_2(t) = t - \ln(1+t) \ .$$

The two following lemmas ensure that the remainder term $\psi_2\left(\left(\frac{s-s_M}{s_M}\right)(z)\right)$ in the expansion of the contrast (17) indeed behaves like a quadratic term, when the unknown density is uniformly bounded from below and elements $s - s_M$ are sufficiently small in sup-norm.

Lemma 6 *Let $\delta \in [0, A_{\min}/2]$. Assume that*

$$\inf_{z \in \mathcal{Z}} s_* (z) \geq A_{\min} > 0 \ . \tag{18}$$

Then, for all $z \in \mathcal{Z}$ and $s \in \widetilde{M}$ such that $|(s - s_M)(z)| \leq \delta$, it holds

$$\left| \left(\frac{s - s_M}{s_M} \right) (z) \right| \leq \frac{\delta}{A_{\min}} \leq \frac{1}{2}$$

and for all $(x, y) \in \left[-\frac{\delta}{A_{\min}}, \frac{\delta}{A_{\min}} \right]$,

$$|\psi_2 (x) - \psi_2 (y)| \leq \frac{2\delta}{A_{\min}} |x - y| \ . \tag{19}$$

Lemma 6 allows us in the Technical Lemmas of Section 5.5 to apply a contraction principle, which can be found in [21] and is recalled in Theorem 28 below, in order to control the second order terms.
Now, the following lemma states that if s is close to s_M in sup-norm, then the Kullback-Leibler divergence is close to a weighted $L_2(P)$ norm.

Lemma 7 *Assume that*

$$\inf_{z \in \mathcal{Z}} s_* (z) \geq A_{\min} > 0 \ . \tag{20}$$

Let $\delta > 0$ such that

$$0 < \delta \leq \frac{A_{\min}}{2} \ .$$

Then for all $s \in M$ such that $\|s - s_M\|_\infty \leq \delta$, we have $\inf_{z \in \mathcal{Z}} s(z) > 0$, and if moreover $\int_{\mathcal{Z}} s d\mu = 1$ then $s \in \widetilde{M}$ and it holds

$$\left(\frac{1}{2} - \frac{2\delta}{3A_{\min}} \right) \left\| \frac{s - s_M}{s_M} \right\|_2^2 \leq \mathcal{K}(s_M, s) = P(Ks - Ks_M) \leq \left(\frac{1}{2} + \frac{2\delta}{3A_{\min}} \right) \left\| \frac{s - s_M}{s_M} \right\|_2^2 \ . \tag{21}$$

The proofs of Lemmas 6 and 7 are postponed to Section 5.1.

3 Results

We state here our main results. In Section 3.1, we investigate the convergence in sup-norm of the histogram estimators towards the Kullback-Leibler projections. This will be needed to derive the sharp upper and lower bounds in probability for the true and empirical excess risks of Section 3.2. Finally, the results obtained in a model selection framework are stated in Section 3.3.

3.1 Rates of convergence in sup-norm of histogram estimators

In order to handle second order terms in the expansion of the contrast (17) we show that the histogram estimator $s_n(M)$ is consistent in sup-norm towards the Kullback-Leibler projection s_M. More precisely, for models having a not too large dimension, the following lemma ensures the convergence in sup-norm of $s_n(M)$ towards s_M at the rate

$$R_{n,D_M} \propto \sqrt{\frac{D_M \ln n}{n}} \ .$$

Proposition 8 *Let α, $A_+, A_*, A_\Lambda > 0$. Consider the linear model M of histograms defined on a finite partition Λ_M of $\mathcal{Z}$, with $|\Lambda_M| = D_M$ its linear dimension. Assume*

$$\|s_*\|_\infty \leq A_* < +\infty \ , \tag{22}$$

$$D_M \inf_{I \in \Lambda_M} \mu(I) \geq A_\Lambda > 0 \ , \tag{23}$$

and

$$D_M \leq A_+ \frac{n}{(\ln n)^2} \leq n \ .$$

Then a positive constant A_c exists, only depending on A_Λ, A_*, A_+ and α such that

$$\mathbb{P}\left[\left\| s_n(M) - s_M \right\|_\infty \geq A_c \sqrt{\frac{D_M \ln n}{n}} \right] \leq 2n^{-\alpha} \ . \tag{24}$$

In Proposition 8, we need to assume that the target s_* is uniformly bounded from above over $\mathcal{Z}$, in order to derive the consistency in sup-norm of the histogram estimator towards the Kullback-Leibler projection s_M. This rather strong assumption can be avoided by normalizing the difference between the histogram estimator and the Kullback-Leibler projection by the latter quantity. The rate of convergence of the sup-norm of the normalized difference is the same as in Proposition 8, that is

$$\sqrt{\frac{D_M \ln n}{n}} \ ,$$

but we assume in Proposition 9 that the target s_* is uniformly bounded away from zero over $\mathcal{Z}$.

Proposition 9 *Let $\alpha,\ A_+, A_{\min}, A_\Lambda > 0$. Consider the linear model M of histograms defined on a finite partition Λ_M of $\mathcal{Z}$, with $|\Lambda_M| = D_M$ its linear dimension. Assume*

$$\inf_{z \in \mathcal{Z}} s_*(z) \geq A_{\min} > +\infty \ , \tag{25}$$

$$D_M \inf_{I \in \Lambda_M} \mu(I) \geq A_\Lambda > 0 \ , \tag{26}$$

and

$$D_M \leq A_+ \frac{n}{(\ln n)^2} \leq n \ .$$

Then a positive constant A_c exists, only depending on A_Λ, $A_{\min}, A_+$ and α such that

$$\mathbb{P}\left[\left\| \frac{s_n(M) - s_M}{s_M} \right\|_\infty \geq A_c \sqrt{\frac{D_M \ln n}{n}} \right] \leq 2n^{-\alpha} \ . \tag{27}$$

As claimed in Remark 11 below, Proposition 9 indeed suffices in the proof of Theorem 10 to handle the second order terms appearing in the expansion of the contrast (17).
The proof of Proposition 8 can be found in Section 5.2.

3.2 True and empirical risks bounds

In this section, we fix the linear model M made of histograms and we are interested by upper and lower bounds for the true excess risk $P(Ks_n(M) - Ks_M)$ on M and for its empirical counterpart $P_n(Ks_M - Ks_n(M))$. We show that under reasonable assumptions the true excess risk is equivalent to the empirical one, which is one of the keystones to prove the slope phenomenon and the optimality of AIC that we state in Section 3.3.

Theorem 10 *Let $\alpha, A_+, A_-, A_{\min}, A_*, A_\Lambda > 0$ and let M be a linear model of histograms defined on a finite partition Λ_M. The finite dimension of M is denoted by D_M. Assume that*

$$0 < A_{\min} \leq \inf_{z \in \mathcal{Z}} s_*(z) \ , \tag{28}$$

$$\|s_*\|_\infty \leq A_* < +\infty \ , \tag{29}$$

$$0 < A_\Lambda \leq D_M \inf_{I \in \Lambda_M} \mu(I) \tag{30}$$

and

$$0 < A_- (\ln n)^2 \leq D_M \leq A_+ \frac{n}{(\ln n)^2} \leq n \ .$$

Then a positive constant A_0 exists, only depending on $\alpha, A_-, A_+, A_*, A_{\min}$ and A_Λ, such that by setting

$$\varepsilon_n(M) = A_0 \max\left\{\left(\frac{\ln n}{D_M}\right)^{1/4}, \left(\frac{D_M \ln n}{n}\right)^{1/4}\right\}, \tag{31}$$

we have, for all $n \geq n_0(A_+, A_-, A_{\min}, A_*, \alpha)$,

$$\mathbb{P}\left[P\left(Ks_n(M) - Ks_M\right) \geq \left(1 - \varepsilon_n(M)\right)\frac{D_M - 1}{2n}\right] \geq 1 - 6n^{-\alpha}, \tag{32}$$

$$\mathbb{P}\left[P\left(Ks_n(M) - Ks_M\right) \leq \left(1 + \varepsilon_n(M)\right)\frac{D_M - 1}{2n}\right] \geq 1 - 6n^{-\alpha}, \tag{33}$$

$$\mathbb{P}\left[P_n\left(Ks_M - Ks_n(M)\right) \geq \left(1 - \varepsilon_n^2(M)\right)\frac{D_M - 1}{2n}\right] \geq 1 - 2n^{-\alpha}, \tag{34}$$

$$\mathbb{P}\left[P_n\left(Ks_M - Ks_n(M)\right) \leq \left(1 + \varepsilon_n^2(M)\right)\frac{D_M - 1}{2n}\right] \geq 1 - 4n^{-\alpha}. \tag{35}$$

In the previous Theorem we achieve sharp upper and lower bounds for the true and empirical excess risk on M. They are optimal at the first order since the leading constants are equal in upper and lower bounds. Moreover, Theorem 10 establishes the equivalence with high probability of the true and empirical excess risks for models of reasonable dimension.
Castellan [14] also asks for a lower regularity property of the partition, for example in Proposition 2.5 where she derive a sharp control of the Kullback-Leibler excess risk of the histogram estimator on a fixed model. More precisely she assumes that there exists a positive constant B such that

$$\inf_{I \in \Lambda_M} \mu(I) \geq B\frac{(\ln n)^2}{n}. \tag{36}$$

This latter assumption is thus weaker than (30) for the considered model as its dimension D_M is less than the order $n(\ln n)^{-2}$. We could assume (36) instead of (30) in order to derive Theorem 10. This would lead to less precise results for second order terms in the deviations of the excess risks but the first order bounds would be preserved. More precisely, if we replace assumption (30) in Theorem 10 by Castellan's assumption (36), a careful look at the proofs of Lemma 4, Proposition 8 and Theorem 10 show that the conclusions of Theorem 10 are still valid for

$$\varepsilon_n = A_0(\ln n)^{-1/4}$$

where A_0 is some positive constant. Thus assumption (30) is not a fundamental restriction in comparison to Castellan's work [14], but it leads to more precise results in terms of deviations of the true and empirical excess risks of the histogram estimator.

Remark 11 *In the proof of Theorem 10 given in Section 5.3 and relying on the technical lemmas given in Section 5.5, we localize the analysis on the subset*

$$B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D_M,\alpha}\right) = \left\{s \in M, \|s - s_M\|_\infty \leq \tilde{R}_{n,D_M,\alpha}\right\},$$

where $\tilde{R}_{n,D_M,\alpha} = A_\infty\sqrt{D_M n^{-1}\ln n}$ is defined in (78). This is possible by using Propostion 8, which states the convergence of $\|s_n(M) - s_M\|_\infty$ towards zero at a rate proportional to $\sqrt{D_M n^{-1}\ln n}$ with high probability. Considering Proposition 9, where we establish the convergence of $\|(s_n(M) - s_M)/s_M\|_\infty$ towards zero, again at a rate proportional to $\sqrt{D_M n^{-1}\ln n}$ with high probability, we can rather localize the analysis on the subset

$$\left\{s \in M, \left\|\frac{s - s_M}{s_M}\right\|_\infty \leq \tilde{R}_{n,D_M,\alpha}\right\}.$$

The gain is that in Proposition 9 - on contrary to Proposition 8 - we do not have to assume that the target s_* is uniformly bounded from above over $\mathcal{Z}$. Hence, a careful look at the proof of Theorem 10, and especially at

the proofs of Lemmas 5, 6 and 7 given in Section 5.1 and the proofs of Lemmas 19, 20, 21 and 22 given in Section 5.5, show that we can make straightforward modifications in order to recover results of Theorem 10 - with different values of the constants - without the assumption (29) of uniform boundedness of the target s_* on $\mathcal{Z}$. More precisely, the other assumptions of Theorem 10 would stay the same, and assumption (29) would be replaced by the much weaker moment condition

$$P\left(\ln s_*\right)_+ < +\infty \ ,$$

ensuring that $s_* \in \mathcal{S}$. The same remark apply to Theorem 12 below.

We turn now to upper bounds in probability for the true and empirical excess risks on models with small dimensions. Our aim here is not to compute sharp constants. In fact, information given by Theorem 12 suffices to our needs as we use it in the proofs of the results stated in Section 3.3 in order to control model selection procedures for small models.

Theorem 12 *Let $\alpha, A_+, A_{\min}, A_*, A_\Lambda > 0$ and let M be a linear model of histograms defined on a finite partition Λ_M. The finite dimension of M is denoted by D_M. Assume that*

$$0 < A_{\min} \leq \inf_{z \in \mathcal{Z}} s_*\left(z\right) \ , \tag{37}$$

$$\left\|s_*\right\|_\infty \leq A_* < +\infty \ , \tag{38}$$

$$0 < A_\Lambda \leq D_M \inf_{I \in \Lambda_M} \mu\left(I\right) \tag{39}$$

and

$$1 \leq D_M \leq A_+ \frac{n}{\left(\ln n\right)^2} \leq n \ .$$

Then a positive constant A_u exists, only depending on $\alpha, A_+, A_*, A_{\min}, A_\Lambda$, such that for all $n \geq n_0\left(A_+, A_*, A_{\min}, A_\Lambda, \alpha\right)$,

$$\mathbb{P}\left[P\left(Ks_n\left(M\right) - Ks_M\right) \geq A_u \frac{D \vee \ln n}{n}\right] \leq 3n^{-\alpha} \ . \tag{40}$$

and

$$\mathbb{P}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right) \geq A_u \frac{D \vee \ln n}{n}\right] \leq 3n^{-\alpha} \ . \tag{41}$$

The proofs of Theorems 10 and 12 can be found in Section 5.3.

3.3 Model Selection

We study in this section the behavior of model selection procedures by penalization of histogram estimators of the density s_*. Under reasonable assumptions stated below, we derive in Theorem 14 a pathwise oracle inequality for the Kullback-Leibler excess risk of the selected estimator, with constant almost one in front of the excess risk of the oracle when the penalty is close to Akaike's one. Our result thus establishes in this case the nonasymptotic quasi-optimality of AIC procedure with respect to the Kullback-Leibler risk. This is an improvement of results of Castellan [14] in the case of "small" collections of models.
Moreover, we validate the slope heuristics first formulated by Birgé and Massart [10] and extended by Arlot and Massart [6]. Indeed, we show in Theorem 13 that if the chosen penalty is less than half of Akaike's penalty then the model selection procedure totally misbehaves. More precisely, the excess risk of the selected estimator is much bigger than the one of the oracle and the dimension of the selected model also explode. This jump of dimension can be exploited in practice to derive a data-driven procedure of calibration of AIC penalty, as explained in Arlot and Massart [6]. This improvement should lead to better performances, at least when the number of data is "small". A comparison, based on simulations, of AIC procedure and the calibration of the linear shape of the optimal penalty via the slope heuristics is still in progress.
Let us now define the model selection procedure. Given a collection of models $\mathcal{M}_n$ with cardinality depending on the number of data n and its associated collection of maximum likelihood estimators

$$\left\{s_n\left(M\right); M \in \mathcal{M}_n\right\} \ ,$$

and a nonnegative penalty function pen on $\mathcal{M}_n$

$$\text{pen} : M \in \mathcal{M}_n \longmapsto \text{pen}(M) \in \mathbb{R}^+$$

the output of the procedure, also called the selected model is

$$\widehat{M} \in \arg\min_{M \in \mathcal{M}_n} \{P_n(Ks_n(M)) + \text{pen}(M)\}. \tag{42}$$

The target of the model selection procedure is

$$M_* \in \arg\min_{M \in \mathcal{M}_n} \{P(Ks_n(M))\}$$

and the associated M-estimator $s_n(M_*)$ is called an oracle. Let us now state the set of assumptions.

3.3.1 Set of assumptions (SA)

(P1) Polynomial complexity of $\mathcal{M}_n$: $\text{Card}(\mathcal{M}_n) \leq c_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}$.

(P2) Upper bound on dimensions of models in $\mathcal{M}_n$: there exists a positive constant $A_{\mathcal{M},+}$ such that for every $M \in \mathcal{M}_n$,

$$D_M \leq A_{\mathcal{M},+} \frac{n}{(\ln n)^2} \leq n \ . \tag{43}$$

(P3) Richness of $\mathcal{M}_n$: there exist $M_0, M_1 \in \mathcal{M}_n$ such that $D_{M_0} \in [\sqrt{n}, c_{rich}\sqrt{n}]$ and $D_{M_1} \geq A_{rich} n (\ln n)^{-2}$.

(Abd) The unknown density s_* is uniformly bounded from below and from above : there exist some positive finite constants $A_{\min}, A_*$ such that,

$$\|s_*\|_\infty \leq A_* < \infty \tag{44}$$

and

$$\inf_{z \in \mathcal{Z}} s_*(z) \geq A_{\min} > 0 \ . \tag{45}$$

($\mathbf{Ap}_u$) The bias decreases as a power of D_M : there exist $\beta_+ > 0$ and $C_+ > 0$ such that

$$\ell(s_*, s_M) \leq C_+ D_M^{-\beta_+} \ .$$

(Alr) Lower regularity of the partition with respect to μ : A positive finite constant A_Λ such that, for all $M \in \mathcal{M}_n$,

$$D_M \inf_{I \in \Lambda_M} \mu(I) \geq A_\Lambda > 0 \ . \tag{46}$$

Theorem 13 *Under the set of assumptions (**SA**) defined above, we further assume that for $A_{\text{pen}} \in [0,1)$ and $A_p > 0$, we have with probability at least $1 - A_p n^{-2}$, for all $M \in \mathcal{M}_n$,*

$$0 \leq \text{pen}(M) \leq A_{\text{pen}} \frac{D_M - 1}{2n} \ . \tag{47}$$

Then there exist two positive constants A_1, A_2 independent of n such that, with probability at least $1 - A_1 n^{-2}$, we have for $n \geq n_0((\boldsymbol{SA}), A_{\text{pen}})$,

$$D_{\widehat{M}} \geq A_2 n \ln(n)^{-2}$$

and

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \geq \ln(n) \inf_{M \in \mathcal{M}_n} \{\ell(s_*, s_n(M))\} \ .$$

In Theorem 13 stated above we prove the existence of a minimal penalty, which is half of AIC. It thus validate the first part of the slope heuristics. Moreover, by Theorem 10 of Section 3.2, we see that for models of dimension not too small we have, with high probability,

$$P_n\left(Ks_M - Ks_n\left(M\right)\right) \approx \frac{D_M - 1}{2n} \ .$$

In fact, a careful look at the proof of Theorem 13 - which follows from arguments that are essentially the same as those of the proof of Theorem 1 of [25] - shows that, by Lemma 16 of Section 5.4, we can replace the condition (47) by the following one,

$$0 \leq \mathrm{pen}\left(M\right) \leq A_{\mathrm{pen}} \mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right] \ .$$

This latter formulation is also interesting because it presents our results as a particular case of the general statement of the slope heuristics given by Arlot and Massart in [6].

Theorem 14 *Assume that the set of assumptions (**SA**) hold together with*

(Ap) *The bias decreases like a power of D_M : there exist $\beta_- \geq \beta_+ > 0$ and $C_+, C_- > 0$ such that*

$$C_- D_M^{-\beta_-} \leq \ell\left(s_*, s_M\right) \leq C_+ D_M^{-\beta_+} \ .$$

Moreover, for $\delta \in \left(0, \frac{1}{2}\right)$ and $L > 0$, assume that an event of probability at least $1 - A_p n^{-2}$ exists on which, for every model $M \in \mathcal{M}_n$ such that $D_M \geq A_{\mathcal{M},+}\left(\ln n\right)^2$,

$$\left(1 - \delta\right) \frac{D_M - 1}{n} \leq \mathrm{pen}\left(M\right) \leq \left(1 + \delta\right) \frac{D_M - 1}{n} \ . \tag{48}$$

Then, for $\frac{1}{2} > \eta > \left(1 - \beta_+\right)_+ / 2$, there exists a constant A_3 and a sequence

$$\theta_n = \sup_{M \in \mathcal{M}_n} \left\{ \varepsilon_n\left(M\right) \ ; \ A_{\mathcal{M},+}\left(\ln n\right)^3 \leq D_M \leq n^{\eta + 1/2} \right\} \leq \frac{L_{(\boldsymbol{SA})}}{\left(\ln n\right)^{1/4}}$$

such that with probability at least $1 - A_3 n^{-2}$, it holds for all $n \geq n_0\left((\boldsymbol{SA}), C_-, \beta_-, \eta, \delta\right)$,

$$D_{\widehat{M}} \leq n^{\eta + 1/2}$$

and

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \leq \left(\frac{1 + 2\delta}{1 - 2\delta} + \frac{5\theta_n}{\left(1 - 2\delta\right)^2}\right) \ell\left(s_*, s_n\left(M_*\right)\right) \ . \tag{49}$$

Theorem 14 states that if the penalty is more than half AIC for models of reasonable dimension then the model selection procedure achieve a nonasymptotic oracle inequality. Moreover, we prove the nonasymptotic quasi-optimality of the selected histogram estimator when the empirical excess risk is penalized by Akaike's criterion, which corresponds to the case where $\delta = 0$. Indeed, we derive in (49) a nonasymptotic pathwise oracle inequality with leading constant almost one. So Theorem 14 validates the second part of the slope heuristics. In order to recover the general formulation of the slope heuristics given by Arlot and Massart, we could replace the condition (48) by the following one

$$2\left(1 - \delta\right) \mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right] \leq \mathrm{pen}\left(M\right) \leq 2\left(1 + \delta\right) \mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right]$$

and the conclusions of the theorem would be exactly the same.
The proofs of Theorems 13 and 14 can be found in Section 5.4.

3.3.2 Comments

Let us now comment on the set of assumptions (**SA**) . Assumption (**P1**) states that the collection of models has a "small" complexity, more precisely a polynomially increasing one. For this kind of complexities, if one wants to perform a good model selection procedure for prediction, the chosen penalty should estimate the mean of the ideal one on each model. Indeed, as Talagrand's type inequalities for the empirical process are pre-Gaussian, they allow to neglect the deviations of the quantities of interest from their mean, uniformly over the collection of models. This is not the case for too large collection of models, where one has to put an extra-log factor depending the complexity of the collection of models inside the penalty, see for example [14] and Massart [23].
The assumption (45) stating that the unknown density is uniformly bounded by below can also be found in the work of Castellan [14]. The author assumes moreover in Theorem 3.4 where she derives an oracle inequality for the Kullback-Leibler excess risk of the histogram estimator, that the target is of finite sup-norm as in inequality (44). But in the case of the Hellinger risk this assumption is replaced in [14] by the weaker assumption that the logarithm of the unknown density s_* is square integrable with respect to the sampling distribution.
In assumption (**P3**) we assume that we have a model M_0 of reasonable dimension and a model M_1 of high dimension. We demand in ($\mathbf{Ap}_u$) that the quality of approximation of the collection of models is good enough in terms of bias. More precisely, we require a polynomially decreasing of excess risk of Kullback-Leibler projections of the unknown density onto the models. For a density s_* uniformly bounded away from zero, this is satisfied when for example, $\mathcal{Z}$ is the unit interval, $\mu = \text{Leb}$ is the Lebesgue measure on the unit interval, the partitions Λ_M are regular and the density s_* belongs to the set $\mathcal{H}(H,\alpha)$ of α-hölderian functions for some $\alpha \in (0,1]$: if $f \in \mathcal{H}(H,\alpha)$, then for all $(x,y) \in \mathcal{Z}^2$

$$|f(x) - f(y)| \leq H |x-y|^{\alpha} \ .$$

In that case, $\beta_+ = 2\alpha$ is convenient and when the chosen penalty is more than half AIC in our case, the procedure is adaptive to the parameters H and α, see Castellan [14].
In assumption (**Ap**) of Theorem 14 we also assume that the bias $\ell(s_*, s_M)$ is lower bounded by a power of the dimension D_M of the model M. This hypothesis is in fact quite classical as it has been used by Stone [26] and Burman [13] for the estimation of density on histograms and also by Arlot and Massart [6] and Arlot [5], [4] in the regression framework. Combining Lemma 1 and 2 of Barron and Sheu [9] we can show that

$$\frac{1}{2} e^{-3\left\| \ln\left(\frac{s_*}{s_M}\right)\right\|_\infty} \int_{\mathcal{Z}} \frac{(s_M - s_*)^2}{s_*} d\mu \leq \ell(s_*, s_M)$$

and thus assuming (**Abd**) we get

$$\frac{A_{\min}^3}{2A_*^4} \int_{\mathcal{Z}} (s_M - s_*)^2 \, d\mu \leq \ell(s_*, s_M) \ .$$

Now, since in the case of histograms the Kullback-leibler projection s_M is also the $L_2(\mu)$ projection of s_* onto M, we can apply Lemma 8.19 in Section 8.10 of Arlot [3] to show that assumption (**Ap**) is satisfied for $\beta_- = 1 + \alpha^{-1}$, in the case where $\mathcal{Z}$ is the unit interval, $\mu = \text{Leb}$ is the Lebesgue measure on the unit interval, the partitions Λ_M are regular and the density s_* is a non-constant α-hölderian function.

4 Two directions of generalization

We present here two possible generalizations of the results exposed in Section 3. Models of piecewise constant densities have the particular property of been exponential models as well as the subset of positive functions in an affine space and we expose below strategies to extend our results in these two directions.
We first notice that the proofs of Theorems 13 and 14 of model selection follow from straightforward adaptations of the proofs of Theorem 2 and 3 in Arlot and Massart [6], only using the results given in Theorems 10, 12 and Lemmas 16 and 17 of Section 5.4 where the quantities of interest can be defined for more general models than histograms. For this reason, the proofs given in Arlot and Massart [6] give some general algebra to derive the properties of the slope heuristics considering a small collection of models and the main task is thus to deal

with some fixed model. Theorems 10 and 12 respectively provide with a sharp control of the excess risk and the empirical excess risk for models of dimension not too large and not too small, and a control of the same quantities for models of small dimension. In Lemma 16 we derive a sharp control of the empirical excess risk in mean for models of reasonable dimension and in Lemma 17 we bound the difference between the bias and its empirical counterpart.
In the following, we emphasize on generalizations of Theorem 10. In fact, Lemma 17 that follows from Bernstein inequality can be easily extended to more general models and Lemma 16 is a straightforward corollary of Theorem 10. Moreover, Theorem 12 directly follows from the convergence in sup-norm of maximum likelihood estimators at the rate $\sqrt{D_M \ln(n)/n}$ as derived in the case of histograms in Proposition 8.

4.1 Affine spaces

We intend to point out here that results of Theorem 10 may be extended to more general linear models M than piecewise constant functions. Let us set

$$M = \left\{ s = \sum_{k=1}^{D_M} \beta_k \varphi_k \ ; \ \beta = (\beta_k)_{k=1}^{D_M} \in \mathbb{R}^{D_M} \right\} \tag{50}$$

the vector space of dimension D_M spanned by the basis $(\varphi_k)_{k=1}^{D_M}$ that we assume to be orthonormal in $L_2(P)$. We also set the subset $\widetilde{M}$ of the functions in M that are densities with respect to μ,

$$\widetilde{M} = \left\{ s \in M \ ; \ s \geq 0 \text{ and } \int_{\mathcal{Z}} s d\mu = 1 \right\} \ ,$$

and consider that the maximum likelihood estimator on $\widetilde{M}$ exists, denoted by $s_n(M)$.
The proof of Theorem 10, that we give in Section 5, relies on purpose on more general arguments than the ones strictly needed in the case of histograms. More precisely, using explicit formula 7 and 10 for the Kullback-Leibler projection and the histogram estimator, we could have avoid the use of the slices in excess risk defined in (87) and (88) by controlling the excess risk and the empirical excess risk directly on the estimator. But our aim is to point out the generality of the method, and a careful look at the proof of Theorem 10 shows that for more general models as in (50), we achieve the same bounds for the excess risks (with different values of constants) if the five following points are satisfied :

- The target s_* is uniformly lower and upper bounded : for $A_{\min}, A_* > 0$,
$$0 < A_{\min} \leq \inf_{z \in \mathcal{Z}} s_*(z) \leq \|s_*\|_\infty \leq A_* < +\infty$$
- The model is of reasonable dimension : $A_- (\ln n)^2 \leq D_M \leq A_+ \frac{n}{(\ln n)^2} \leq n$.
- $(\varphi_k)_{k=1}^{D_M}$ is a localized orthonormal basis in $(M, L_2(P))$: for some $r_M > 0$,
$$\left\| \sum_{k=1}^{D_M} \beta_k \varphi_k \right\|_\infty \leq r_M \sqrt{D_M} |\beta|_\infty \ . \tag{51}$$
- The Kullback-Leibler projection s_M is well-defined and the excess risk is, locally around s_M, close to the weighted $L_2(P)$ norm : positive constants A_H and L_H exist such that, if $\|s - s_M\|_\infty \leq \delta \leq A_H$ then
$$\left(\frac{1}{2} - L_H \delta\right) \left\| \frac{s - s_M}{s_M} \right\|_2^2 \leq P(Ks - Ks_M) \leq \left(\frac{1}{2} + L_H \delta\right) \left\| \frac{s - s_M}{s_M} \right\|_2^2 \ . \tag{52}$$
- The maximum likelihood estimator is consistent towards the Kullback-Leibler projection s_M at the rate $\sqrt{D_M \ln(n)/n}$: for any $\alpha > 0$, positive constants A_c and L_c exist such that
$$\mathbb{P}\left[\|s_n(M) - s_M\|_\infty \geq A_c \sqrt{\frac{D_M \ln n}{n}} \right] \leq L_c n^{-\alpha} \ . \tag{53}$$

Note that the assumption of lower regularity of the partition of Theorem 10 in the case of histograms, stating that $D_M \inf_{I \in \Lambda_M} \mu(I) \geq A_\Lambda > 0$ for some $A_\Lambda > 0$, is replaced here by the more general assumption of localized basis (51). It is easy to see using Lemma 4 that the two properties are equivalent in the case of histograms. Moreover, Property (52) is based in the case of histograms on the Pythagorean-like identity (9) given in Proposition 1 and remains a work in progress for more general models $\widetilde{M}$. In Csiszár and Matúš [17], general conditions are given under which Pythagorean-like identities for the Kullback-Leibler divergence hold true. In their terminology, the Kullback-Leibler projection is called "reverse I-projection". Among other results, they show Pythagorean-like identities in the context of convex sets, a property that is satisfied for $\widetilde{M}$, but considering the "I-projection" rather than the "reverse I-projection". Nevertheless, generalized reverse I-projections onto convex sets of probability measures can be found in Barron [7]. Property (53) remains an open issue for general linear models as well.

4.2 Exponential models

In this section, we briefly describe how our strategy of proofs can be adapted to derive sharp bounds for the excess risks in the case of exponential models and possibly recover the slope heuristics in good cases. This work is still in progress. Let us set

$$M = \left\{ t = \sum_{k=1}^{D_M} \beta_k \varphi_k \ ; \ \beta = (\beta_k)_{k=1}^{D_M} \in \mathbb{R}^{D_M} \right\}$$

the linear vector space of dimension D_M spanned by the basis $(\varphi_k)_{k=1}^{D_M}$, that we assume to be orthonormal in $L_2(P)$. We assume that the constant function $\mathbf{1} \in M$ and that $M \subset L_\infty(\mu)$. Then we set the associated exponential model $\widetilde{M}$, defined to be

$$\widetilde{M} = \left\{ s = \exp(t) \ ; t \in M \text{ and } \int_{\mathcal{Z}} s d\mu = 1 \right\}$$

and consider the maximum likelihood estimator $s_n\left(\widetilde{M}\right)$ on $\widetilde{M}$. It is well-known (see for example Barron and Sheu [9] and also Csiszár and Matúš [17]) that in this case $s_n(M)$ exists with high probability as a solution of a family of linear constraints, and its uniqueness is a familiar consequence of the strict convexity of the log-likelihood. It is also well-known (see Lemma 3 of Barron and Sheu [9]) that the unknown density s_* has a unique Kullback-Leibler projection $s_{\widetilde{M}}$on $\widetilde{M}$, characterized by the following Pythagorean-like identity,

$$\mathcal{K}(s_*, s) = \mathcal{K}\left(s_*, s_{\widetilde{M}}\right) + \mathcal{K}\left(s_{\widetilde{M}}, s\right) \ .$$

This property is essential, as it follows that the excess risk on $\widetilde{M}$ is the Kullback-Leibler divergence with respect to the Kullback-Leibler projection $s_{\widetilde{M}}$,

$$P\left(Ks - Ks_{\widetilde{M}}\right) = \mathcal{K}\left(s_{\widetilde{M}}, s\right)$$

and by consequence, we can relate the excess risk on $\widetilde{M}$ to the $L_2(P)$ norm in M, due to the following lemma of Barron and Sheu [9].

Lemma 15 (Lemma 3, [9]) *Let p and q be two probability density functions with respect to μ such that $\|\ln(p/q)\|_\infty$ is finite. Then it holds*

$$\mathcal{K}(p, q) \geq \frac{1}{2} e^{-\|\ln(p/q)\|_\infty} \int p \left(\ln \frac{p}{q} \right)^2 d\mu$$

and

$$\mathcal{K}(p, q) \leq \frac{1}{2} e^{\|\ln(p/q) - c\|_\infty} \int p \left(\ln \frac{p}{q} - c \right)^2 d\mu \ ,$$

where c is any constant.

Hence, we have for any $s \in \widetilde{M}$,

$$0 < \frac{1}{2} e^{-\left\| \ln\left(s/s_{\widetilde{M}}\right)\right\|_\infty} \int s_{\widetilde{M}} \left(\ln \frac{s}{s_{\widetilde{M}}} \right)^2 d\mu \leq P\left(Ks - Ks_{\widetilde{M}}\right) \tag{54}$$

$$\leq \frac{1}{2} e^{\left\| \ln\left(s/s_{\widetilde{M}}\right)\right\|_\infty} \int s_{\widetilde{M}} \left(\ln \frac{s}{s_{\widetilde{M}}} \right)^2 d\mu \ . \tag{55}$$

Now, if we can show that

$$\left\| \ln \left(\frac{s_n\left(\widetilde{M}\right)}{s_{\widetilde{M}}} \right) \right\|_\infty \leq \frac{A_{cons}}{\sqrt{\ln n}}$$

for some positive constant A_{cons} and for all n sufficiently large, we can restrict our study to the subset of functions in $\widetilde{M}$ satisfying $\left\| \ln\left(s/s_{\widetilde{M}}\right)\right\|_\infty \leq \frac{A_{cons}}{\sqrt{\ln n}}$ - by the same type of arguments that are given in Section 6 of [24] - and so we have on this subset of interest, by inequalities (54) and (55),

$$P\left(Ks - Ks_{\widetilde{M}}\right) \sim \frac{1}{2} \int s_{\widetilde{M}} \left(\ln \frac{s}{s_{\widetilde{M}}} \right)^2 d\mu$$

$$= \frac{1}{2} \left\| \ln \frac{s}{s_{\widetilde{M}}} \right\|_2^2 + \frac{1}{2} \int \left(s_{\widetilde{M}} - s_*\right) \left(\ln \frac{s}{s_{\widetilde{M}}} \right)^2 d\mu \ .$$

Moreover, for the right-hand term in the latter identity, it holds

$$\left| \frac{1}{2} \int \left(s_{\widetilde{M}} - s_*\right) \left(\ln \frac{s}{s_{\widetilde{M}}} \right)^2 d\mu \right| \leq \left\| s_{\widetilde{M}} - s_* \right\|_\infty \frac{1}{2} \left\| \ln \frac{s}{s_{\widetilde{M}}} \right\|_{L_2(\mu)}^2$$

which should be negligible in front of the weighted $L_2(P)$ norm $\frac{1}{2} \left\| \ln \frac{s}{s_{\widetilde{M}}} \right\|_2^2$ if the considered model $\widetilde{M}$ has a small bias in sup-norm and if the unknown density is uniformly bounded away from zero, in order to upper bound $\|\cdot\|_{L_2(\mu)}$ by $\|\cdot\|_{L_2(P)}$. Under the right assumptions on the smoothness of the target s_* and a suitable choice of M the assumption on the bias of the model should be satisfied if at least its dimension is not too small (a power of $\ln n$ should be again sufficient in many cases). The importance of a control in sup-norm for the bias of the models in maximum likelihood estimation of density has been pointed out by Stone [27] considering log-splines models. The author provides with a sharp control of the bias in sup-norm in this case, a work that should be inspiring for other situations and also in order to prove the consistency in sup-norm of $\ln\left(\frac{s_n\left(\widetilde{M}\right)}{s_{\widetilde{M}}}\right)$. By consequence, we can conjecture that under reasonable assumptions, the weighted $L_2(P)$ norm described above is a good approximation of the excess risk on $\widetilde{M}$ for a model M of dimension not too small and it has the convenient property to be Hilbertian : on a subset of interest on $\widetilde{M}$,

$$P\left(Ks - Ks_{\widetilde{M}}\right) \sim \frac{1}{2} \left\| \ln \frac{s}{s_{\widetilde{M}}} \right\|_2^2 = \frac{1}{2} \left\| \ln s - \ln s_{\widetilde{M}} \right\|_2^2 \tag{56}$$

where $\ln(s)$ and $\ln\left(s_{\widetilde{M}}\right)$ belong to M.

Let us explain now how to take advantage of (56) for exponential models. The arguments given below are close in the spirit to arguments of Section 6 of [24], considering the log-linearity of exponential models, or in other words the linearity of the contrasted functions. If we set

$$t_M = \ln s_{\widetilde{M}} \in M$$

and for any $r \geq 0$,

$$\xi_n(r) = \mathbb{E}\left[\sup_{\substack{t \in M,\ \|t - t_M\|_2^2 = 2r \\ \int \exp(t) d\mu = 1}} \left| (P - P_n)(t - t_M) \right| \right] \ ,$$

then, as claimed in Section 6 of [24], we can approximately write for models of reasonable dimension,

$$P\left(Ks_n\left(\widetilde{M}\right)-Ks_M\right)\sim \arg\max_{R_{n,D_M}\geq r\geq 0}\left\{\mathbb{E}\left[\sup_{s\in\widetilde{M},\ P\left(Ks-Ks_{\widetilde{M}}\right)=r}\left|\left(P-P_n\right)\left(Ks-Ks_{\widetilde{M}}\right)\right|\right]-r\right\}$$

where we assume that

$$P\left(Ks_n\left(\widetilde{M}\right)-Ks_{\widetilde{M}}\right)\leq\left\|\ln\left(\frac{s_n\left(\widetilde{M}\right)}{s_{\widetilde{M}}}\right)\right\|_\infty\leq R_{n,D_M}\leq\frac{A_{cons}}{\sqrt{\ln n}}$$

with high probability (of order $1-Ln^{-\alpha}$, $\alpha>0$). Then, from (56) we have for $R_{n,D_M}\geq r\geq 0$,

$$\mathbb{E}\left[\sup_{s\in\widetilde{M},\ P\left(Ks-Ks_{\widetilde{M}}\right)=r}\left|\left(P-P_n\right)\left(Ks-Ks_{\widetilde{M}}\right)\right|\right]\sim\xi_n\left(r\right)$$

and so

$$P\left(Ks_n\left(\widetilde{M}\right)-Ks_M\right)\sim\arg\max_{R_{n,D_M}\geq r\geq 0}\left\{\xi_n\left(r\right)-r\right\}\ . \tag{57}$$

By the same type of reasoning, we can also conjecture that for models of reasonable dimensions,

$$P_n\left(Ks_M-Ks_n\left(\widetilde{M}\right)\right)\sim\max_{R_{n,D_M}\geq r\geq 0}\left\{\xi_n\left(r\right)-r\right\}\ . \tag{58}$$

Moreover, in good cases satisfying assumptions of Corollary 32 we have

$$\xi_n\left(r\right)\sim\mathbb{E}^{1/2}\left[\left(\sup_{\substack{t\in M,\ \|t-t_M\|_2^2=2r\\ \int\exp(t)d\mu=1}}\left|\left(P-P_n\right)\left(t-t_M\right)\right|\right)^2\right] \tag{59}$$

and if we define

$$t_{CS}=\sqrt{2r}\frac{\sum_{k=1}^{D_M}\left(P-P_n\right)\left(\varphi_k\right)\varphi_k}{\sqrt{\sum_{k=1}^{D_M}\left(P-P_n\right)^2\left(\varphi_k\right)}}+t_M\ ,$$

then it holds $\|t_{CS}-t_M\|_2^2=2r$ and

$$\begin{aligned}\sup_{t\in M,\ \|t-t_M\|_2^2=2r}\left|\left(P-P_n\right)\left(t-t_M\right)\right|&=\left(P-P_n\right)\left(t_{CS}-t_M\right)\\&=\sqrt{2r}\sqrt{\sum_{k=1}^{D_M}\left(P-P_n\right)^2\left(\varphi_k\right)}\ .\end{aligned} \tag{60}$$

Now, assuming that $1\gg R_{n,D_M}\geq L\sqrt{\frac{D_M\ln n}{n}}$ for a positive constant L sufficiently large, if we can prove that with high probability,

$$\|t_{CS}-t_M\|_\infty\leq R_{n,D_M}\text{ for }r\leq R_{n,D_M}\ ,$$

which is typically the case when $(\varphi_k)_{k=1}^{D_M}$ is a localized basis, then

$$\begin{aligned}\int\exp(t_{CS})d\mu&\approx\int\exp(t_M)d\mu+\int\left(t_{CS}-t_M\right)d\mu\\&\approx 1\ .\end{aligned} \tag{61}$$

Finally, taking into account (59), (60) and (61), we can conjecture that under some assumptions on the model M that allow to control the sup-norm in a sufficiently sharp way, we would have

$$\xi_n(r) \sim \sqrt{\frac{2r}{n}\sum_{k=1}^{D_M}\mathrm{Var}(\varphi_k)}$$

for $r \leq R_{n,D_M}$ and so, using (57) and (58), as

$$\arg\max_{R_{n,D_M}\geq r\geq 0}\left\{\sqrt{\frac{2r}{n}\sum_{k=1}^{D_M}\mathrm{Var}(\varphi_k)}-r\right\} = \max_{R_{n,D_M}\geq r\geq 0}\left\{\sqrt{\frac{2r}{n}\sum_{k=1}^{D_M}\mathrm{Var}(\varphi_k)}-r\right\} = \frac{\sum_{k=1}^{D_M}\mathrm{Var}(\varphi_k)}{2n}$$

for $R_{n,D_M} \geq \sqrt{\frac{\sum_{k=1}^{D_M}\mathrm{Var}(\varphi_k)}{2n}}$, this would lead to

$$P\left(Ks_n\left(\widetilde{M}\right)-Ks_M\right) \sim P_n\left(Ks_M - Ks_n\left(\widetilde{M}\right)\right) \sim \frac{\sum_{k=1}^{D_M}\mathrm{Var}(\varphi_k)}{2n}$$

for models of reasonable dimensions having good enough properties with respect to the sup-norm.

5 Proofs

5.1 Proofs of Section 2

Proof of Lemma 4. Remind that, for all $I \in \Lambda_M$,

$$\varphi_I = (P(I))^{-1/2}\mathbf{1}_I\ .$$

Hence, $(\varphi_I)_{I\in\Lambda_M}$ is an orthonormal basis of $(M, L_2(P))$. Moreover, by (12) we have, for all $I \in \Lambda_M$,

$$P(I) \geq A_{\min}\mu(I) \geq A_{\min}A_\Lambda D_M^{-1} > 0$$

and so, by setting $r_M = (A_{\min}A_\Lambda)^{-1/2}$, we get for all $I \in \Lambda_M$,

$$(P(I))^{-1/2} \leq \sqrt{\frac{D_M}{A_{\min}A_\Lambda}} = r_M\sqrt{D_M}\ .$$

Now, as the elements φ_I for $I \in \Lambda_M$ have disjoint supports, we deduce that, for all $\beta = (\beta_I)_{I\in\Lambda_M} \in \mathbb{R}^{D_M}$,

$$\begin{aligned}\left\|\sum_{I\in\Lambda_M}\beta_I\varphi_I\right\|_\infty &\leq \max_{I\in\Lambda_M}\{|\beta_I|\,\|\varphi_I\|_\infty\}\\ &\leq \max_{I\in\Lambda_M}\left\{|\beta_I|\,(P(I))^{-1/2}\right\}\\ &\leq r_M\sqrt{D_M}\,|\beta|_\infty\end{aligned}$$

and Inequality (13) is then proved. Next, Inequality (14) easy follows by observing that, for any $s = \sum_{I\in\Lambda_M}\beta_I\varphi_I \in M$ satisfying $\|s\|_2 \leq 1$, we have

$$\max_{I\in\Lambda_M}|\beta_I| \leq \sqrt{\sum_{I\in\Lambda_M}\beta_I^2} \leq 1$$

and so

$$\|s\|_\infty \leq r_M\sqrt{D_M}\ .$$

■

Proof of Lemma 5. By (15)and (7), we have

$$\inf_{z\in\mathcal{Z}} s_M(z) \geq A_{\min} > 0 \ ,$$

then $\psi_{1,M}(z)$ and $(Ks_M)(z) = -\ln(s_M(z))$ are well defined for all $z \in \mathcal{Z}$. Moreover, as we assume $\|s - s_M\|_\infty < A_{\min}$, we have

$$\inf_{z\in\mathcal{Z}} s(z) = \inf_{z\in\mathcal{Z}} \{s_M(z) + (s - s_M)(z)\} \geq \inf_{z\in\mathcal{Z}} s_M(z) - \|s - s_M\|_\infty > 0$$

and

$$\left\| \frac{s - s_M}{s_M} \right\|_\infty \leq \frac{\|s - s_M\|_\infty}{A_{\min}} < 1$$

thus $(Ks)(z) = -\ln(s(z))$ is well defined for each $z \in \mathcal{Z}$ as well as $(s_M(z))^{-1}$ and $\ln\left(1 + \frac{s-s_M}{s_M}(z)\right)$, so the expansion (17) is a simple rewriting of the identity

$$(Ks)(z) - (Ks_M)(z) = -\ln\left(\frac{s(z)}{s_M(z)}\right) \ .$$

■

Proof of Lemma 6. Lemma **6** is straightforward, since

$$\psi_2'(x) = \frac{x}{1+x},\ x \in (-1, +\infty) \ .$$

Hence, for all $x \in \left[-\frac{\delta}{A_{\min}}, \frac{\delta}{A_{\min}}\right]$, with $0 \leq \delta \leq A_{\min}/2$,

$$|h'(x)| \leq \frac{\delta/A_{\min}}{1 - \delta/A_{\min}} \leq 2\frac{\delta}{A_{\min}} \ ,$$

which yields the result. ■

Proof of Lemma 7. For $s \in M$ such that $\|s - s_M\|_\infty \leq \delta \leq \frac{A_{\min}}{2}$, we have

$$\inf_{z\in\mathcal{Z}} s(z) \geq \inf_{z\in\mathcal{Z}} s_M(z) - \|s - s_M\|_\infty \geq \frac{A_{\min}}{2} > 0 \text{ and } \left\| \frac{s - s_M}{s_M} \right\|_\infty \leq \frac{1}{2} \ .$$

and so, if $\int_{\mathcal{Z}} s d\mu = 1$ then $s \in \widetilde{M}$. Moreover, in this case, by (11) we have

$$P(Ks - Ks_M) = \mathcal{K}(s_M, s)$$

and it holds

$$\begin{aligned}
\mathcal{K}(s_M, s) &= \int_{\mathcal{Z}} \ln\left(\frac{s_M}{s}\right) s_M d\mu \\
&= \int_{\mathcal{Z}} -\ln\left(1 + \frac{s - s_M}{s_M}\right) s_M d\mu \\
&= \sum_{k=1}^{\infty} \frac{(-1)^k}{k} \int_{\mathcal{Z}} \left(\frac{s - s_M}{s_M}\right)^k s_M d\mu \\
&= \int_{\mathcal{Z}} (s_M - s) d\mu + \frac{1}{2} \int_{\mathcal{Z}} \left(\frac{s - s_M}{s_M}\right)^2 s_M d\mu + \sum_{k=3}^{\infty} \frac{(-1)^k}{k} \int_{\mathcal{Z}} \left(\frac{s - s_M}{s_M}\right)^k s_M d\mu. \qquad (62)
\end{aligned}$$

Now, as $\int_{\mathcal{Z}} s d\mu = 1$, we have

$$\int_{\mathcal{Z}} (s_M - s)\, d\mu = 0 \ . \tag{63}$$

Moreover, notice that by (7), for all $I \in \Lambda_M$,

$$\int_{\mathcal{Z}} \mathbf{1}_I s_M d\mu = \frac{P(I)}{\mu(I)} \mu(I) = P(I) = \int_{\mathcal{Z}} \mathbf{1}_I s_* d\mu$$

and so, for all $t \in M$,

$$\int_{\mathcal{Z}} t \cdot s_M d\mu = \int_{\mathcal{Z}} t \cdot s_* d\mu \ .$$

Now, using the fact that $\left(\frac{s - s_M}{s_M}\right)^2 \in M$, it holds

$$\begin{aligned} \frac{1}{2} \int_{\mathcal{Z}} \left(\frac{s - s_M}{s_M}\right)^2 s_M d\mu &= \frac{1}{2} \int_{\mathcal{Z}} \left(\frac{s - s_M}{s_M}\right)^2 s_* d\mu \\ &= \frac{1}{2} P \left(\frac{s - s_M}{s_M}\right)^2 \\ &= \frac{1}{2} \left\| \frac{s - s_M}{s_M} \right\|_2^2 \ . \end{aligned} \tag{64}$$

Moreover, we have

$$\begin{aligned} &\left| \sum_{k=3}^{\infty} \frac{(-1)^k}{k} \int_{\mathcal{Z}} \left(\frac{s - s_M}{s_M}\right)^k s_M d\mu \right| \\ &\leq \frac{1}{3} \int_{\mathcal{Z}} \left(\frac{s - s_M}{s_M}\right)^2 s_M d\mu \times \sum_{j=1}^{\infty} \left\| \frac{s - s_M}{s_M} \right\|_{\infty}^j \\ &= \frac{1}{3} \int_{\mathcal{Z}} \left(\frac{s - s_M}{s_M}\right)^2 s_* d\mu \times \sum_{j=1}^{\infty} \left\| \frac{s - s_M}{s_M} \right\|_{\infty}^j \\ &\leq \left\| \frac{s - s_M}{s_M} \right\|_2^2 \frac{2\delta}{3A_{\min}} \ . \end{aligned} \tag{65}$$

Inequality (21) then follows by using (63), (64) and (65) in (62). ■

5.2 Proof of Section 3.1

Proof of Proposition 8. Let $\beta > 0$ to be fixed later. Recall that, by (7) and (10),

$$\begin{aligned} s_n(M) &= \sum_{I \in \Lambda_M} \frac{P_n(I)}{\mu(I)} \mathbf{1}_I \ , \\ s_M &= \sum_{I \in \Lambda_M} \frac{P(I)}{\mu(I)} \mathbf{1}_I \ . \end{aligned}$$

Hence, the sup-norm of the difference can be written

$$\|s_n(M) - s_M\|_{\infty} = \sup_{I \in \Lambda_M} \frac{|(P_n - P)(I)|}{\mu(I)} \ . \tag{66}$$

By Bernstein's inequality (171) applied for the random variable $\mathbf{1}_{\xi\in I}$ we get, for all $x>0$,

$$\mathbb{P}\left[|(P_n-P)(I)|\geq\sqrt{\frac{2P(I)\,x}{n}}+\frac{x}{3n}\right]\leq 2\exp(-x)\ .$$

Taking $x=\beta\ln n$ and normalizing by the quantity $\mu(I)>0$ we get

$$\mathbb{P}\left[\frac{|(P_n-P)(I)|}{\mu(I)}\geq\frac{1}{\mu(I)}\sqrt{\frac{2\beta P(I)\ln n}{n}}+\frac{\beta\ln n}{\mu(I)\,3n}\right]\leq 2n^{-\beta}\ . \tag{67}$$

Now, by (22) and (23),

$$0<\frac{1}{\mu(I)}\leq\frac{D_M}{A_\Lambda} \tag{68}$$

and

$$\frac{\sqrt{P(I)}}{\mu(I)}\leq\sqrt{\frac{A_*}{\mu(I)}}\leq\sqrt{\frac{A_*D_M}{A_\Lambda}}\ . \tag{69}$$

So, injecting (68) and (69) in (67) and using the fact that $D_M\leq A_+\frac{n}{(\ln n)^2}$ we get

$$\mathbb{P}\left[\frac{|(P_n-P)(I)|}{\mu(I)}\geq A_c\sqrt{\frac{D_M\ln n}{n}}\right]\leq 2n^{-\beta}\ , \tag{70}$$

where $A_c=\max\left\{\sqrt{\frac{2\beta A_*}{A_\Lambda}}\ ;\ \frac{\beta\sqrt{A_+}}{3A_\Lambda}\right\}$. We then deduce from (66) and (70) that

$$\mathbb{P}\left[\|s_n(M)-s_M\|_\infty\geq A_c\sqrt{\frac{D_M\ln n}{n}}\right]\leq\frac{2D_M}{n^\beta}$$

and, since $D_M\leq n$, taking $\beta=\alpha+1$ yields Inequality (24). ■

Proof of Proposition 9. Let $\beta>0$ to be fixed later. Recall that, by (7) and (10),

$$s_n(M)=\sum_{I\in\Lambda_M}\frac{P_n(I)}{\mu(I)}\mathbf{1}_I\ , \tag{71}$$

$$s_M=\sum_{I\in\Lambda_M}\frac{P(I)}{\mu(I)}\mathbf{1}_I\ . \tag{72}$$

Hence, by (25) and (72) we get $\inf s_M(z)\geq A_{\min}>0$. By (71) and (72) we have

$$\left\|\frac{s_n(M)-s_M}{s_M}\right\|_\infty=\sup_{I\in\Lambda_M}\frac{|(P_n-P)(I)|}{P(I)}\ . \tag{73}$$

By Bernstein's inequality (171) applied for the random variable $\mathbf{1}_{\xi\in I}$ we get, for all $x>0$,

$$\mathbb{P}\left[|(P_n-P)(I)|\geq\sqrt{\frac{2P(I)\,x}{n}}+\frac{x}{3n}\right]\leq 2\exp(-x)\ .$$

Taking $x=\beta\ln n$ and normalizing by the quantity $P(I)\geq A_{\min}\mu(I)>0$ we get

$$\mathbb{P}\left[\frac{|(P_n-P)(I)|}{P(I)}\geq\sqrt{\frac{2\beta\ln n}{P(I)\,n}}+\frac{\beta\ln n}{P(I)\,3n}\right]\leq 2n^{-\beta}\ . \tag{74}$$

Now, by (25) and (26), we have

$$0 < \frac{1}{P(I)} \leq \frac{D_M}{A_{\min} A_\Lambda} \ . \tag{75}$$

Hence, using (75) in (74) and using the fact that $D_M \leq A_+ \frac{n}{(\ln n)^2}$ we get

$$\mathbb{P}\left[\frac{|(P_n - P)(I)|}{P(I)} \geq A_c \sqrt{\frac{D_M \ln n}{n}}\right] \leq 2n^{-\beta} \ , \tag{76}$$

where $A_c = \max\left\{\sqrt{\frac{2\beta}{A_\Lambda A_{\min}}} \ ; \ \frac{\beta\sqrt{A_+}}{3A_{\min}A_\Lambda}\right\}$. We then deduce from (73) and (76) that

$$\mathbb{P}\left[\|s_n(M) - s_M\|_\infty \geq A_c \sqrt{\frac{D_M \ln n}{n}}\right] \leq \frac{2D_M}{n^\beta}$$

and, since $D_M \leq n$, taking $\beta = \alpha + 1$ yields Inequality (27). ■

5.3 Proofs of Theorems 10 and 12

In order to introduce the quantities of interest, we recall some notations stated below and add some new definitions. As usual, M denotes the finite dimensional linear vector space of piecewise constant functions with respect to the finite partition Λ_M. Moreover, we write $D_M = |\Lambda_M|$ the linear dimension of M. Assuming (46) and (45) we have, for all $I \in \Lambda_M$, $P(I) > 0$ and so, if we set

$$\varphi_I = \frac{\mathbf{1}_I}{\sqrt{P(I)}} \ , \qquad I \in \Lambda_M \ ,$$

the family $(\varphi_I)_{I \in \Lambda_M}$ is an orthonormal basis of $(M, L_2(P))$. We set

$$\nu_n = \max\left\{\sqrt{\frac{\ln n}{D_M}}, \ \sqrt{\frac{D_M \ln n}{n}}\right\} \ . \tag{77}$$

In what follows $\alpha > 0$ is fixed and for some positive constant A_∞ to be chosen in the proof of Theorem 10 and satisfying

$$A_\infty \geq A_c > 0$$

where A_c is defined in Proposition 8 and only depends on A_Λ, A_*, A_+ and α, we set

$$\tilde{R}_{n,D_M,\alpha} = A_\infty \sqrt{\frac{D_M \ln n}{n}} \tag{78}$$

and

$$\Omega_{\infty,\alpha} = \left\{\|s_n(M) - s_M\|_\infty \leq \tilde{R}_{n,D_M,\alpha}\right\} \ .$$

By Proposition 8 it holds, since $A_\infty \geq A_c$,

$$\mathbb{P}\left[\Omega^c_{\infty,\alpha}\right] \leq 2n^{-\alpha} \ . \tag{79}$$

Moreover, our analysis is localized on the subset

$$B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D_M,\alpha}\right) = \left\{s \in M, \|s - s_M\|_\infty \leq \tilde{R}_{n,D_M,\alpha}\right\} \ .$$

Assuming that

$$D_M \leq A_+ n (\ln n)^{-2}$$

we have, for all $n \geq n_0(A_+, A_{\min}, A_\infty)$,

$$\tilde{R}_{n,D_M,\alpha} \leq \frac{A_{\min}}{2} \tag{80}$$

where $A_{\min}$ is defined in (45). Now, assuming (45), we have by (80) and Lemma 5, for all $n \geq n_0\left(A_+, A_{\min}, A_\infty\right)$, for every $s \in B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D_M,\alpha}\right)$ and all $z \in \mathcal{Z}$,

$$(Ks)(z) - (Ks_M)(z) = \psi_{1,M}(z)(s - s_M)(z) + \psi_2\left(\left(\frac{s - s_M}{s_M}\right)(z)\right) \tag{81}$$

where

$$\psi_{1,M}(z) = -\frac{1}{s_M(z)}$$

and, for all $t \in (-1, +\infty)$,

$$\psi_2(t) = t - \ln(1+t) \ .$$

Recall that, by (45),

$$\left\|\psi_{1,M}\right\|_\infty \leq \left(\inf_{z\in\mathcal{Z}} |s_M(z)|\right)^{-1} \leq A_{\min}^{-1} \ . \tag{82}$$

Moreover, by (80) and Lemma 6 we have, for all $n \geq n_0\left(A_+, A_{\min}, A_\infty\right)$, for all $s \in B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D_M,\alpha}\right)$ and all $z \in \mathcal{Z}$, using that $\psi_2(0) = 0$,

$$\left|\psi_2\left(\left(\frac{s - s_M}{s_M}\right)(z)\right)\right| \leq \left|\left(\frac{s - s_M}{s_M}\right)(z)\right| \ . \tag{83}$$

We also have by (80) and Lemma 6, for all $n \geq n_0\left(A_+, A_{\min}, A_\infty\right)$, for every $s, t \in B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D_M,\alpha}\right)$ and all $z \in \mathcal{Z}$,

$$\left|\psi_2\left(\left(\frac{s - s_M}{s_M}\right)(z)\right) - \psi_2\left(\left(\frac{t - s_M}{s_M}\right)(z)\right)\right| \leq 2A_{\min}^{-2}\tilde{R}_{n,D_M,\alpha}\left|(t-s)(z)\right| \ . \tag{84}$$

For convenience, we will use the following notation,

$$\psi_2 \circ \left(\frac{s - s_M}{s_M}\right) : z \in \mathcal{Z} \longmapsto \psi_2\left(\left(\frac{s - s_M}{s_M}\right)(z)\right) \ .$$

We now define slices of excess risk on the model $\widetilde{M}$. We set, for all $C > 0$,

$$\mathcal{F}_C = \left\{s \in \widetilde{M} \ ; \ \left\|\psi_{1,M} \cdot (s - s_M)\right\|_2^2 \leq 2C\right\} \bigcap B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D_M,\alpha}\right) \tag{85}$$

$$\mathcal{F}_{>C} = \left\{s \in \widetilde{M} \ ; \ \left\|\psi_{1,M} \cdot (s - s_M)\right\|_2^2 > 2C\right\} \bigcap B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D_M,\alpha}\right) \tag{86}$$

and for any interval J,

$$\mathcal{F}_J = \left\{s \in \widetilde{M} \ ; \ \frac{1}{2}\left\|\psi_{1,M} \cdot (s - s_M)\right\|_2^2 \in J\right\} \bigcap B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D_M,\alpha}\right) \ . \tag{87}$$

We also define, for all $L \geq 0$,

$$D_L = \left\{s \in \widetilde{M} \ ; \ \left\|\psi_{1,M} \cdot (s - s_M)\right\|_2^2 = 2L\right\} \bigcap B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D_M,\alpha}\right) \ . \tag{88}$$

By Lemma 7, we have, for all $n \geq n_0\left(A_+, A_{\min}, A_\infty\right)$ and for any $s \in B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D_M,\alpha}\right)$ such that $\int_{\mathcal{Z}} s d\mu = 1$,

$$0 < \left(\frac{1}{2} - \frac{2}{3A_{\min}}\tilde{R}_{n,D_M,\alpha}\right)\left\|\psi_{1,M} \cdot (s - s_M)\right\|_2^2 \leq \mathcal{K}(s_M, s) = P(Ks - Ks_M) \tag{89}$$

$$\leq \left(\frac{1}{2} + \frac{2}{3A_{\min}}\tilde{R}_{n,D_M,\alpha}\right)\left\|\psi_{1,M} \cdot (s - s_M)\right\|_2^2 \ . \tag{90}$$

Finally, notice that, if we assume (45) and (**Alr**), then by Proposition 4, if we set $r_M = (A_{\min} A_\Lambda)^{-1/2}$ then for all $z \in \mathcal{Z}$,

$$\sup_{s \in M,\ \|s\|_2 \leq 1} \|s\|_\infty \leq r_M \sqrt{D_M} \tag{91}$$

and moreover, for all $\beta = (\beta_I)_{I \in \Lambda_M} \in \mathbb{R}^{D_M}$,

$$\left\| \sum_{I \in \Lambda_M} \beta_I \varphi_I \right\|_\infty \leq r_M \sqrt{D_M}\, |\beta|_\infty \ . \tag{92}$$

5.3.1 Proofs of Theorems 10 and 12.

Proof of Theorem (10). We divide the proof of Theorem 10 in four parts corresponding to the four Inequalities (32), (33), (34) and (35). The values of A_0 and A_∞, respectively defined in (31) and (78), will then be fixed at the end of the proof. Note that, since $D_M \geq A_- (\ln n)^2$, we have $D_M \geq 2$ for all $n \geq n_0(A_-)$ so we can assume in the following that $D_M \geq 2$.

Proof of Inequality (32). By (78), it holds for all $n \geq n_0(A_+, A_{\min}, A_\infty)$,

$$1 - \frac{4}{3A_{\min}} \tilde{R}_{n, D_M, \alpha} > \frac{1}{2} \ .$$

Let $r \in (1, 2]$ to be chosen later and $C, \widetilde{C} > 0$ such that

$$rC = \frac{D_M - 1}{2n} \tag{93}$$

and, for all $n \geq n_0(A_+, A_{\min}, A_\infty)$,

$$\widetilde{C} = \left(1 - \frac{4}{3A_{\min}} \tilde{R}_{n, D_M, \alpha}\right) C > 0 \ .$$

By inequality (89), if

$$P(Ks_n(M) - Ks_M) \leq \widetilde{C} \quad \text{and} \quad \|s_n(M) - s_M\|_\infty \leq \tilde{R}_{n, D_M, \alpha}$$

then

$$\left\|\psi_{1,M} \cdot (s_n(M) - s_M)\right\|_2^2 \leq 2C \ ,$$

for all $n \geq n_0(A_+, A_{\min}, A_\infty)$. Hence, by inequality (79), we get for all $n \geq n_0(A_+, A_{\min}, A_\infty)$,

$$\begin{aligned}
\mathbb{P}\left(P(Ks_n(M) - Ks_M) \leq \widetilde{C}\right) &\leq \mathbb{P}\left(\left\{P(Ks_n(M) - Ks_M) \leq \widetilde{C}\right\} \bigcap \Omega_{\infty,\alpha}\right) + 2n^{-\alpha} \\
&\leq \mathbb{P}\left(\left\{\left\|\psi_{1,M} \cdot (s_n(M) - s_M)\right\|_2^2 \leq 2C\right\} \bigcap \Omega_{\infty,\alpha}\right) + 2n^{-\alpha} \ .
\end{aligned} \tag{94}$$

Now, by definition of the slices $\mathcal{F}_C$ and $\mathcal{F}_{>C}$ respectively given in (85) and (86), it holds

$$\begin{aligned}
&\mathbb{P}\left(\left\{\left\|\psi_{1,M} \cdot (s_n(M) - s_M)\right\|_2^2 \leq 2C\right\} \bigcap \Omega_{\infty,\alpha}\right) \\
&\leq \mathbb{P}\left(\inf_{s \in \mathcal{F}_C} P_n(Ks - Ks_M) \leq \inf_{s \in \mathcal{F}_{>C}} P_n(Ks - Ks_M)\right) \\
&\leq \mathbb{P}\left(\inf_{s \in \mathcal{F}_C} P_n(Ks - Ks_M) \leq \inf_{s \in \mathcal{F}_{(C, rC]}} P_n(Ks - Ks_M)\right) \\
&= \mathbb{P}\left(\sup_{s \in \mathcal{F}_C} P_n(Ks_M - Ks) \geq \sup_{s \in \mathcal{F}_{(C, rC]}} P_n(Ks_M - Ks)\right) \ .
\end{aligned} \tag{95}$$

Now, as by (93) we have

$$\frac{D_M}{8n} \leq C \leq (1 + A_4\nu_n)^2 \frac{D_M - 1}{2n}$$

where A_4 is defined in Lemma 23, we can apply Lemma 23 with $\alpha = \beta$, $A_l = 1/8$ and it holds, for all $n \geq n_0(A_-, A_+, A_{\min}, r_M, A_\infty, \alpha)$,

$$\mathbb{P}\left[\sup_{s \in \mathcal{F}_C} P_n(Ks_M - Ks) \geq \left(1 + L_{A_-, A_{\min}, A_\infty, r_M, \alpha} \times \nu_n\right) \sqrt{\frac{2C(D_M - 1)}{n}} - C\right] \leq 2n^{-\alpha} , \tag{96}$$

where $\nu_n = \max\left\{\sqrt{\frac{\ln n}{D_M}}, \sqrt{\frac{D_M \ln n}{n}}\right\}$. Moreover, we can apply Lemma 25 with

$$\alpha = \beta, \ A_l = 1/8, \ A_u = 1/2$$

and

$$A_\infty \geq 32\sqrt{2} B_2 A_* r_M ,$$

and since $rC = (D_M - 1)/2n$, it gives, for all $n \geq n_0(A_+, A_-, r_M, A_{\min}, A_\infty)$,

$$\mathbb{P}\left(\sup_{s \in \mathcal{F}_{(C, rC]}} P_n(Ks_M - Ks) \leq \left(\frac{1}{2} - L_{A_-, A_{\min}, A_\infty, \alpha} \times \nu_n\right) \frac{D_M - 1}{n}\right) \leq 2n^{-\alpha} , \tag{97}$$

Now, from (96) and (97) we can deduce that a positive constant $\tilde{A}_0$ exists, only depending on A_-, $A_{\min}$, A_∞, r_M and α, such that for all $n \geq n_0(A_-, A_+, A_{\min}, r_M, A_\infty, \alpha)$, it holds on the same event of probability at least $1 - 4n^{-\alpha}$,

$$\begin{aligned}\sup_{s \in \mathcal{F}_C} P_n(Ks_M - Ks) &\leq \left(1 + \tilde{A}_0\nu_n\right) \sqrt{\frac{2C(D_M - 1)}{n}} - C \\ &= \left(1 + \tilde{A}_0\nu_n\right) \frac{D_M - 1}{n} \frac{1}{\sqrt{r}} - \frac{D_M - 1}{2n} \frac{1}{r}\end{aligned} \tag{98}$$

and

$$\sup_{s \in \mathcal{F}_{(C, rC]}} P_n(Ks_M - Ks) \geq \left(1 - 2\tilde{A}_0\nu_n\right) \frac{D_M - 1}{2n} . \tag{99}$$

Hence, from (98) and (99) we can deduce, using (94) and (95), that if we choose $r \in (1, 2]$ such that

$$\left(1 - 2\tilde{A}_0\nu_n\right) r - 2\left(1 + \tilde{A}_0\nu_n\right)\sqrt{r} + 1 > 0 \tag{100}$$

then, for all $n \geq n_0(A_-, A_+, A_{\min}, r_M, A_\infty, \alpha)$, $P(Ks_n(M) - Ks_M) \geq C$ with probability at least $1 - 6n^{-\alpha}$. Moreover, since

$$A_-(\ln n)^2 \leq D_M \leq A_+ n (\ln n)^{-2}$$

we have, for all $n \geq n_0\left(A_+, A_-, \tilde{A}_0\right)$,

$$\tilde{A}_0\nu_n \leq \frac{1}{4} \tag{101}$$

and so, for all $n \geq n_0\left(A_+, A_-, \tilde{A}_0\right)$, simple computations using (101) show that by taking

$$r = 1 + 48\sqrt{\tilde{A}_0\nu_n} \tag{102}$$

inequality (100) is satisfied. Notice that, for all $n \geq n_0\left(A_+, A_-, \tilde{A}_0\right)$, $0 < 48\sqrt{\tilde{A}_0\nu_n} < 1$, so that $r \in (1, 2]$. Finally, we can compute C by (93) and (102), for all $n \geq n_0\left(A_+, A_-, \tilde{A}_0\right)$,

$$C = \frac{rC}{r} = \frac{1}{1 + 48\sqrt{\tilde{A}_0\nu_n}} \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \geq \left(1 - 48\sqrt{\tilde{A}_0\nu_n}\right) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 > 0 . \tag{103}$$

The result then follows the fact that by (103) and (77), it holds for all $n \geq n_0\left(A_+, A_-, A_\infty, A_{\min}, \tilde{A}_0\right)$,

$$\begin{aligned}
\widetilde{C} &= \left(1 - \frac{4}{3A_{\min}}\tilde{R}_{n,D_M,\alpha}\right) C \\
&\geq \left(1 - \frac{4}{3A_{\min}}\tilde{R}_{n,D_M,\alpha}\right)\left(1 - 48\sqrt{\tilde{A}_0\nu_n}\right)\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2 \\
&\geq \left(1 - \frac{4}{3A_{\min}}\tilde{R}_{n,D_M,\alpha}\right)\left(1 - 48\sqrt{\tilde{A}_0\nu_n}\right)\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2 \\
&\geq \left(1 - L_{A_\infty,A_{\min}}\nu_n\right)\left(1 - 48\sqrt{\tilde{A}_0\nu_n}\right)\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2 \\
&\geq \left(1 - L_{A_\infty,A_{\min},\tilde{A}_0}\sqrt{\nu_n}\right)\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2 \ ,
\end{aligned}$$

where the constant $\tilde{A}_0$ only depends on A_-, $A_{\min}$, A_∞, r_M and α. ■

To prove inequalities (33), (34), (35) and Theorem 12 it suffices to adapt the proofs of inequalities (23), (24), (25) and Theorem 4 given in Section 7 of [24] in the same way that we just did in the proof of inequality (32). We thus skip these proofs as they are now straightforward.

5.4 Proofs of Section 3.3

Given Lemmas 16 and 17 below, the proofs of Theorems 13 and 14 follow from straightforward adaptations of the proofs of Theorems 1 and 2 given in Section 4 of [25].

Lemma 16 *Let $A_{\mathcal{M},-} > 0$. Assume (**P2**), (**Abd**) and (**Alr**) of the set of assumptions defined in Section 3.3.1. Then for every model M of dimension D_M such that*

$$A_{\mathcal{M},-}\left(\ln n\right)^2 \leq D_M \leq A_{\mathcal{M},+} n \left(\ln n\right)^{-2} \ ,$$

we have, for all $n \geq n_0\left(A_{\mathcal{M},+}, A_{\mathcal{M},-}, A_\Lambda, A_{\min}, A_*, \alpha_{\mathcal{M}}\right)$,

$$\left(1 - L_{A_{\mathcal{M},+},A_{\mathcal{M},-},A_*,A_{\min},A_\Lambda}\varepsilon_n^2\left(M\right)\right)\frac{D_M - 1}{2n} \leq \mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right] \tag{104}$$

$$\leq \left(1 + L_{A_{\mathcal{M},+},A_{\mathcal{M},-},A_*,A_{\min},A_\Lambda}\varepsilon_n^2\left(M\right)\right)\frac{D_M - 1}{2n} \tag{105}$$

where $\varepsilon_n\left(M\right) = A_0 \max\left\{\left(\frac{\ln n}{D_M}\right)^{1/4}, \left(\frac{D_M \ln n}{n}\right)^{1/4}\right\}$ is defined in Theorem 10.

Proof. Under assumptions of Lemma 16 we can apply Theorem 10 with $\alpha = 2 + \alpha_{\mathcal{M}}$. For all $n \geq n_0\left(A_{\mathcal{M},+}, A_{\mathcal{M},-}, A_{\min}, A_*, \alpha_{\mathcal{M}}\right)$, we thus have on an event $\Omega_1\left(M\right)$ of probability at least $1 - 6n^{-2-\alpha_{\mathcal{M}}}$,

$$\left(1 - \varepsilon_n^2\left(M\right)\right)\frac{D_M - 1}{2n} \leq P_n\left(Ks_M - Ks_n\left(M\right)\right) \leq \left(1 + \varepsilon_n^2\left(M\right)\right)\frac{D_M - 1}{2n} \tag{106}$$

where

$$\varepsilon_n\left(M\right) = A_0 \max\left\{\left(\frac{\ln n}{D_M}\right)^{1/4}, \left(\frac{D_M \ln n}{n}\right)^{1/4}\right\} \geq A_0 n^{-1/8} \tag{107}$$

as $D_M \geq 1$. Moreover, we have,

$$\begin{aligned}
0 &\leq P_n\left(Ks_M - Ks_n(M)\right) = P_n\left(\ln\left(\frac{s_n(M)}{s_M}\right)\right) \\
&= P_n\left(\sum_{I\in\Lambda_M} \ln\left(\frac{P_n(I)}{P(I)}\right)\mathbf{1}_I\right) = \sum_{I\in\Lambda_M} \ln\left(P_n(I)\right) P_n(I) + \sum_{I\in\Lambda_M} \ln\left(\frac{1}{P(I)}\right) P_n(I) \\
&\leq \max_{I\in\Lambda_M}\left\{\ln\left(\frac{1}{P(I)}\right)\right\} \leq \ln\left(\left(A_{\min}A_\Lambda\right)^{-1} D_M\right) \ , \qquad (108)
\end{aligned}$$

where the last inequality follows from (**Abd**) and (**Alr**). We also have

$$\begin{aligned}
&\mathbb{E}\left[P_n\left(Ks_M - Ks_n(M)\right)\right] \\
&= \mathbb{E}\left[P_n\left(Ks_M - Ks_n(M)\right)\mathbf{1}_{\Omega_1(M)}\right] + \mathbb{E}\left[P_n\left(Ks_M - Ks_n(M)\right)\mathbf{1}_{(\Omega_1(M))^c}\right]. \qquad (109)
\end{aligned}$$

Hence, as $n \geq D_M \geq A_{\mathcal{M},-}\left(\ln n\right)^2$, it comes from (107) and (108) that, for all $n \geq n_0\left(A_{\mathcal{M},-}, A_0, A_{\min}, A_\Lambda\right)$,

$$0 \leq \mathbb{E}\left[P_n\left(Ks_M - Ks_n(M)\right)\mathbf{1}_{(\Omega_1(M))^c}\right] \leq 6\ln\left(\left(A_{\min}A_\Lambda\right)^{-1} D_M\right) n^{-2-\alpha_{\mathcal{M}}} \leq \varepsilon_n^2(M)\,\frac{D_M - 1}{2n} \qquad (110)$$

and, as we can see that $\varepsilon_n(M) < 1$ for all $n \geq n_0(A_0)$, we have by (106), for all $n \geq n_0\left(A_{\mathcal{M},+}, A_{\mathcal{M},-}, A_0, A_{\min}, A_*, \alpha_{\mathcal{M}}\right)$,

$$\begin{aligned}
\left(1 - 6n^{-2-\alpha_{\mathcal{M}}}\right)\left(1 - \varepsilon_n^2(M)\right)\frac{D_M - 1}{2n} &\leq \mathbb{E}\left[P_n\left(Ks_M - Ks_n(M)\right)\mathbf{1}_{\Omega_1(M)}\right] \qquad (111) \\
&\leq \left(1 - 6n^{-2-\alpha_{\mathcal{M}}}\right)\left(1 + \varepsilon_n^2(M)\right)\frac{D_M - 1}{2n} \ . \qquad (112)
\end{aligned}$$

Finally, noticing that $n^{-2-\alpha_{\mathcal{M}}} \leq A_0^{-2}\varepsilon_n^2(M)$ by (107), we can use (110), (111) and (112) in (109) to conclude by straightforward computations that

$$L_{A_{\mathcal{M},+},A_{\mathcal{M},-},A_*,A_{\min},A_\Lambda} = 6A_0^{-2} + 2$$

is convenient in (104) and (105), as A_0 only depends on $\alpha_{\mathcal{M}}, A_-, A_+, A_*, A_{\min}$ and A_Λ. ■

Lemma 17 *Let $\alpha > 0$. Assume that (**Abd**) of Section 3.3.1 is satisfied. Then by setting $\bar{\delta}(M) = (P_n - P)(Ks_M - Ks_*)$, we have for all $M \in \mathcal{M}_n$,*

$$\mathbb{P}\left(\left|\bar{\delta}(M)\right| \geq \sqrt{\frac{4A_*\alpha\ell(s_*, s_M)\ln n}{A_{\min}n}} + \ln\left(\frac{A_*}{A_{\min}}\right)\frac{\alpha\ln n}{3n}\right) \leq 2n^{-\alpha} \qquad (113)$$

*and if moreover, assumptions (**P2**), (**Abd**) and (**Alr**) of Section 3.3.1 hold, then a positive constant A_d exists, depending only in A_*, $A_{\min}$ and α such that, for all $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},-}\left(\ln n\right)^2 \leq D_M$ and for all $n \geq n_0\left(A_{\mathcal{M},+}, A_{\mathcal{M},-}, A_*, A_{\min}, A_\Lambda\right)$,*

$$\mathbb{P}\left(\left|\bar{\delta}(M)\right| \geq \frac{\ell(s_*, s_M)}{\sqrt{D_M}} + A_d\frac{\ln n}{\sqrt{D_M}}\mathbb{E}\left[p_2(M)\right]\right) \leq 2n^{-\alpha} \ , \qquad (114)$$

where $p_2(M) = P_n\left(Ks_M - Ks_n(M)\right)$.

Proof. First recall that

$$s_M = \sum_{I\in\Lambda_M}\frac{P(I)}{\mu(I)}\mathbf{1}_I = \sum_{I\in\Lambda_M}\left(\int_I s_*\frac{d\mu}{\mu(I)}\right)\mathbf{1}_I \ .$$

Thus by (**Abd**) we deduce that

$$0 < A_{\min} \leq \inf_{z \in \mathcal{Z}} s_M(z) \leq \|s_M\|_\infty \leq A_* < +\infty \ . \tag{115}$$

Now, as we have

$$Ks_M - Ks_* = -\ln\left(\frac{s_M}{s_*}\right) \ ,$$

we get, by (**Abd**) and (115), that

$$\|Ks_M - Ks_*\|_\infty \leq \ln\left(\frac{A_*}{A_{\min}}\right) \ . \tag{116}$$

Hence, by Lemma 1 of Barron and Sheu [9], we have

$$P\left[\left(Ks_M - Ks_*\right)^2\right] \leq 2\exp\left(\|Ks_M - Ks_*\|_\infty\right)\mathcal{K}(s_*, s_M) \ .$$

By Proposition 1, we also have

$$\mathcal{K}(s_*, s_M) = P(Ks_M - Ks_*) = \ell(s_*, s_M)$$

and thus by (116), it holds

$$P\left[\left(Ks_M - Ks_*\right)^2\right] \leq \frac{2A_*}{A_{\min}}\ell(s_*, s_M) \ . \tag{117}$$

We are now ready to apply Bernstein's inequality (171) to

$$\bar{\delta}(M) = (P_n - P)(Ks_M - Ks_*) \ .$$

By (116) and (117) we have, for any $x > 0$,

$$\mathbb{P}\left(\left|\bar{\delta}(M)\right| \geq \sqrt{\frac{4A_*\ell(s_*, s_M)\,x}{A_{\min}n}} + \ln\left(\frac{A_*}{A_{\min}}\right)\frac{x}{3n}\right) \leq 2\exp(-x) \ .$$

Hence, taking $x = \alpha \ln n$ we have

$$\mathbb{P}\left(\left|\bar{\delta}(M)\right| \geq \sqrt{\frac{4A_*\alpha\ell(s_*, s_M)\ln n}{A_{\min}n}} + \ln\left(\frac{A_*}{A_{\min}}\right)\frac{\alpha\ln n}{3n}\right) \leq 2n^{-\alpha} \ , \tag{118}$$

which yields Inequality (113). Now, by noticing the fact that $2\sqrt{ab} \leq a\eta + b\eta^{-1}$ for all $\eta > 0$, and by using it in (118) with $a = \ell(s_*, s_M)$, $b = \frac{A_*\alpha\ln n}{A_{\min}n}$ and $\eta = D_M^{-1/2}$, we obtain

$$\mathbb{P}\left(\left|\bar{\delta}(M)\right| \geq \frac{\ell(s_*, s_M)}{\sqrt{D_M}} + \left(\frac{A_*}{A_{\min}}\sqrt{D_M} + \frac{1}{3}\ln\left(\frac{A_*}{A_{\min}}\right)\right)\frac{\alpha\ln n}{n}\right) \leq 2n^{-\alpha} \ . \tag{119}$$

Then, for a model M such that $A_{\mathcal{M},-}(\ln n)^2 \leq D_M \leq A_{\mathcal{M},+}n(\ln n)^{-2}$, we can apply Lemma 16 and by (104), it holds for all $n \geq n_0(A_{\mathcal{M},+}, A_{\mathcal{M},-}, A_\Lambda, A_{\min}, A_*, \alpha_{\mathcal{M}})$,

$$\left(1 - L_{A_{\mathcal{M},+}, A_{\mathcal{M},-}, A_*, A_{\min}, A_\Lambda}\varepsilon_n^2(M)\right)\frac{D_M - 1}{2n} \leq \mathbb{E}\left[p_2(M)\right] \tag{120}$$

where $\varepsilon_n(M) = A_0 \max\left\{\left(\frac{\ln n}{D_M}\right)^{1/4}, \left(\frac{D_M \ln n}{n}\right)^{1/4}\right\}$. Moreover as

$$A_{\mathcal{M},-}(\ln n)^2 \leq D_M \leq A_{\mathcal{M},+}n(\ln n)^{-2} \ ,$$

we can deduce that for all $n \geq n_0(A_{\mathcal{M},+}, A_{\mathcal{M},-}, A_*, A_{\min}, A_\Lambda)$,

$$L_{A_{\mathcal{M},+}, A_{\mathcal{M},-}, A_*, A_{\min}, A_\Lambda}\varepsilon_n^2(M) \leq 1/2$$

and we have by (120), $\mathbb{E}\left[p_2(M)\right] \geq \frac{D_M}{8n}$ for all $n \geq n_0(A_{\mathcal{M},+}, A_{\mathcal{M},-}, A_*, A_{\min}, A_\Lambda)$. This allows, using (119), to conclude the proof by simple computations. ■

5.5 Technical lemmas

We state here some lemmas needed in the proofs of Theorem 10. Their proofs are quite similar to the proofs given in Section 7 of [24] as we use the same generic approach exposed in details in Section 6 of [24]. More precisely, the least-squares contrast in regression and the Kullback-Leibler contrast satisfy the same formal property of expansion (81) and the models that we consider are endowed with localized basis. The main technical difference comes from the fact that the Kullback-Leibler excess risk is only close to an Hilbertian norm on the considered functions of $B_{(M,L_\infty)}(s_M, \tilde{R}_{n,D_M,\alpha})$, whereas in the least-squares regression the excess risk is the Hilbertian $L_2(P)$ norm itself.

Lemma 18 *Assume (45), (**Alr**) and $D_M \geq 2$. Then for any $\beta > 0$, a positive constant $L_{r_M,\beta}$ exists, such that by setting*

$$\tau_n = L_{r_M,\beta}\left(\sqrt{\frac{\ln n}{D_M}} \vee \frac{\sqrt{\ln n}}{n^{1/4}}\right) ,$$

we have

$$\mathbb{P}\left[\sqrt{\sum_{I\in\Lambda_M}(P_n - P)^2(\varphi_I)} \geq (1+\tau_n)\sqrt{\frac{D_M - 1}{n}}\right] \leq n^{-\beta} .$$

Proof. By Cauchy-Schwarz inequality we have

$$\chi_M = \sqrt{\sum_{I\in\Lambda_M}(P_n - P)^2(\varphi_I)} = \sup_{s\in\mathcal{F}_{(C,rC]}}\{|(P_n - P)(s)| \ ; \ s\in M \ \& \ \|s\|_2 \leq 1\} .$$

Hence, we get by Bousquet's inequality (173) with $\mathcal{F} = \{s \ ; \ s\in M, \ \|s\|_2 \leq 1\}$, for all $x > 0$, $\delta > 0$,

$$\mathbb{P}\left[\chi_M \geq \sqrt{2\sigma^2\frac{x}{n}} + (1+\delta)\,\mathbb{E}[\chi_M] + \left(\frac{1}{3} + \frac{1}{\delta}\right)\frac{bx}{n}\right] \leq \exp(-x) \tag{121}$$

where

$$\sigma^2 \leq \sup_{s\in M, \ \|s\|_2\leq 1}\mathbb{E}\left[(s(X))^2\right] = 1$$

and

$$b \leq \sup_{s\in M, \ \|s\|_2\leq 1}\|s - P(s)\|_\infty \leq 2\sup_{s\in M, \ \|s\|_2\leq 1}\|s\|_\infty \leq 2r_M\sqrt{D_M} \qquad \text{by (91).}$$

Moreover, since

$$\sum_{I\in\Lambda_M}\mathrm{Var}(\varphi_I) = \sum_{I\in\Lambda_M}(1 - P(I)) = D_M - 1 ,$$

it holds

$$\mathbb{E}[\chi_M] \leq \sqrt{\mathbb{E}[\chi_M^2]} = \sqrt{\frac{\sum_{I\in\Lambda_M}\mathrm{Var}(\varphi_I)}{n}} = \sqrt{\frac{D_M - 1}{n}} .$$

So, from (121) it follows that

$$\mathbb{P}\left[\chi_M \geq \sqrt{\frac{2x}{n}} + (1+\delta)\sqrt{\frac{D_M - 1}{n}} + \left(\frac{1}{3} + \frac{1}{\delta}\right)\frac{2r_M\sqrt{D_M}x}{n}\right] \leq \exp(-x) . \tag{122}$$

Hence, taking $x = \beta\ln n$, $\delta = \frac{\sqrt{\ln n}}{n^{1/4}}$ in (122), we can derive that a positive constant $L_{r_M,\beta}$ exists such that

$$\mathbb{P}\left[\chi_M \geq \left(1 + L_{r_M,\beta}\left(\sqrt{\frac{\ln n}{D_M}} \vee \frac{\sqrt{\ln n}}{n^{1/4}}\right)\right)\sqrt{\frac{D_M - 1}{n}}\right] \leq n^{-\beta} ,$$

which gives the result.■

Lemma 19 *Let $r > 1$ and $C > 0$. Assume that (**Abd**) and (**Alr**) hold. If positive constants A_-, A_+, A_l, A_u exist such that*

$$A_+ \frac{n}{(\ln n)^2} \geq D \geq A_- (\ln n)^2 \quad \text{and} \quad A_l \frac{D}{n} \leq rC \leq A_u \frac{D}{n} \ ,$$

and if the constant A_∞ defined in (78) satisfies

$$A_\infty \geq 64 B_2 \sqrt{A_u} A_* r_M \ , \tag{123}$$

then a positive constant $L_{A_l, A_u, A_{\min}}$ exists such that, for all $n \geq n_0 (B_2, A_+, A_-, A_l, A_u, r_M, A_{\min})$,

$$\mathbb{E}\left[\sup_{s \in \mathcal{F}_{(C,rC]}} (P_n - P)\left(\psi_{1,M} \cdot (s_M - s)\right) \right] \geq \left(1 - \frac{L_{A_l, A_u, A_{\min}}}{\sqrt{D_M}}\right) \sqrt{\frac{2rC(D_M - 1)}{n}} \ . \tag{124}$$

In the previous Lemma, we state a sharp lower bound for the mean of the supremum of the empirical process on the linear parts of contrasted functions of $\widetilde{M}$ belonging to a slice of excess risk. This is done for models of reasonable dimensions. Moreover, we see that we need to assume that the constant A_∞ introduced in (78) is large enough. In order to prove Lemma 19 we need the following intermediate result.

Lemma 20 *Let $r > 1$, $A_+, A_-, A_u, \beta > 0$ and $C \geq 0$. Assume that (**Abd**) and (**Alr**) hold and that*

$$A_+ \frac{n}{(\ln n)^2} \geq D_M \geq A_- (\ln n)^2 \quad \text{and} \quad rC \leq A_u \frac{D_M}{n} \ .$$

Set

$$\beta_{n,I} = \frac{\sqrt{2rC}\,(P_n - P)(\varphi_I)}{\sqrt{\sum_{I \in \Lambda_M} (P_n - P)^2 (\varphi_I)}} \qquad \text{for all } I \in \Lambda_M \ ,$$

and

$$s_{CS} = \sum_{I \in \Lambda_M} \beta_{n,I} \varphi_I \ \in M \ .$$

Then the following inequality holds,

$$\int_{\mathcal{Z}} (s_M s_{CS} + s_M)\, d\mu = 1 \tag{125}$$

and if the constant A_∞ defined in (78) satisfies

$$A_\infty \geq 32 B_2 \sqrt{A_u \beta} A_* r_M \ ,$$

then it holds, for all $n \geq n_0 (B_2, A_+, A_-, r_M, \beta)$„

$$\mathbb{P}\left[\max_{I \in \Lambda_M} |\beta_{n,I}| \geq \frac{\tilde{R}_{n, D_M, \alpha}}{A_* r_M \sqrt{D_M}} \right] \leq \frac{2D_M + 1}{n^\beta} \ . \tag{126}$$

In this case, $(s_M \times s_{CS} + s_M) \in \mathcal{F}_{(C,rC]}$ with probability at least $1 - (2D_M + 1) n^{-\beta}$.

Proof of Lemma 20. Let us begin with property (125). As $\int_{\mathcal{Z}} s_M d\mu = 1$, it suffices to check that

$$\int_{\mathcal{Z}} s_M \times s_{CS} d\mu = 0 \ .$$

Indeed, as by (7) we have $s_M = \sum_{I\in\Lambda_M} \frac{P(I)}{\mu(I)}\mathbf{1}_I$,

$$\begin{aligned} s_M \times s_{CS} &= \frac{\sqrt{2rC}}{\sqrt{\sum_{I\in\Lambda_M}(P_n-P)^2(\varphi_I)}} \sum_{I\in\Lambda_M}(P_n-P)\left(\frac{\mathbf{1}_I}{\sqrt{P(I)}}\right)\frac{P(I)}{\mu(I)}\frac{\mathbf{1}_I}{\sqrt{P(I)}} \\ &= \frac{\sqrt{2rC}}{\sqrt{\sum_{I\in\Lambda_M}(P_n-P)^2(\varphi_I)}} \sum_{I\in\Lambda_M}(P_n-P)(\mathbf{1}_I)\frac{\mathbf{1}_I}{\mu(I)} \ . \end{aligned}$$

So the expectation of $s_M \times s_{CS}$ with respect to μ is proportional to

$$\begin{aligned} &\int_{\mathcal{Z}} \sum_{I\in\Lambda_M}(P_n-P)(\mathbf{1}_I)\frac{\mathbf{1}_I}{\mu(I)}d\mu \\ &= (P_n-P)(1_{\mathcal{Z}}) = 0 \ . \end{aligned}$$

Thus property (125) is satisfied. We now turn to the proof of (126). As in the proof of Lemma 18 we write

$$\chi_M = \sqrt{\sum_{I\in\Lambda_M}(P_n-P)^2(\varphi_I)} \ .$$

By Cauchy-Schwarz inequality, we get

$$\chi = \sup_{s\in S_M} |(P_n-P)(s)| \ ,$$

where S_M is the unit sphere of M, that is

$$S_M = \left\{ s\in M, s = \sum_{I\in\Lambda_M}\beta_I\varphi_I \text{ and } \sqrt{\sum_{I\in\Lambda_M}\beta_I^2} = 1 \right\} \ .$$

Thus we can apply Klein-Rio's bound (175) to χ since it holds

$$\begin{aligned} &\sup_{s\in S_M}\|s-Ps\|_\infty \le 2\sup_{s\in S_M}\|s\|_\infty \le 2r_M\sqrt{D_M} \qquad \text{by (91).} \\ &\sup_{s\in S_M}\mathrm{Var}(s) \le 1 \end{aligned} \tag{127}$$

and also, by Inequality (170), using (127),

$$\begin{aligned} \mathbb{E}[\chi_M] &\ge B_2^{-1}\sqrt{\mathbb{E}[\chi_M^2]} - \frac{2r_M\sqrt{D_M}}{n} \\ &= B_2^{-1}\sqrt{\frac{D_M-1}{n}} - \frac{2r_M\sqrt{D_M}}{n} \ . \end{aligned}$$

We thus obtain, for all $\varepsilon, x > 0$,

$$\mathbb{P}\left[\chi_M \le (1-\varepsilon)B_2^{-1}\sqrt{\frac{D_M-1}{n}} - \sqrt{2\frac{x}{n}} - \left(1-\varepsilon+\left(1+\frac{1}{\varepsilon}\right)x\right)\frac{2r_M\sqrt{D_M}}{n}\right] \le \exp(-x) \ .$$

So, by taking $\varepsilon = \frac{1}{2}$ and $x = \beta\ln n$, and as $D_M \ge A_-(\ln n)^2$, it holds, for all $n \ge n_0(B_2, A_-, r_M, \beta)$,

$$\mathbb{P}\left[\chi_M \le \frac{B_2^{-1}}{8}\sqrt{\frac{D_M}{n}}\right] \le n^{-\beta} \ . \tag{128}$$

Furthermore, combining Bernstein's inequality (171) with the observation that we have, for every $I \in \Lambda_M$,

$$\begin{aligned} \|\varphi_I - P\varphi_I\|_\infty \leq 2\|\varphi_I\|_\infty \leq 2r_M\sqrt{D_M} \qquad & \text{by (92)} \\ \mathrm{Var}(\varphi_I) \leq 1 \ , & \end{aligned}$$

we get that, for every $x > 0$,

$$\mathbb{P}\left[|(P_n - P)(\varphi_I)| \geq \sqrt{2\frac{x}{n}} + \frac{2r_M\sqrt{D_M}x}{3n}\right] \leq 2\exp(-x) \ .$$

Hence, for $x = \beta \ln n$, it comes

$$\mathbb{P}\left[\max_{I\in\Lambda_M} |(P_n - P)(\varphi_I)| \geq \sqrt{\frac{2\beta\ln n}{n}} + \frac{2r_M\sqrt{D_M}\beta\ln n}{3n}\right] \leq \frac{2D_M}{n^\beta} \ , \tag{129}$$

then by using (128) and (129), for all $n \geq n_0(B_2, A_-, r_M, \beta)$,

$$\mathbb{P}\left[\max_{I\in\Lambda_M} |\beta_{n,I}| \geq \frac{8B_2\sqrt{2rC}}{\sqrt{\frac{D_M}{n}}}\left(\sqrt{\frac{2\beta\ln n}{n}} + \frac{2r_M\sqrt{D_M}\beta\ln n}{3n}\right)\right] \leq \frac{2D_M+1}{n^\beta} \ .$$

Finally, as $A_+\frac{n}{(\ln n)^2} \geq D_M$ we have, for all $n \geq n_0(A_+, r_M, \beta)$,

$$\frac{2r_M\sqrt{D_M}\beta\ln n}{3n} \leq \sqrt{\frac{2\beta\ln n}{n}}$$

and we can check that if

$$A_\infty \geq 32B_2\sqrt{A_u\beta}A_* r_M$$

then, for all $n \geq n_0(B_2, A_+, A_-, r_M, \beta)$,

$$\mathbb{P}\left[\max_{I\in\Lambda_M} |\beta_{n,I}| \geq \frac{A_\infty}{A_* r_M}\sqrt{\frac{\ln n}{n}}\right] \leq \frac{2D_M+1}{n^\beta} \ .$$

which readily yields Inequality (126). As a consequence, it holds with probability at least $1 - (2D_M+1)n^{-\beta}$,

$$\begin{aligned} \|(s_M \times s_{CS} + s_M) - s_M\|_\infty &\leq \|s_M\|_\infty \|s_{CS}\|_\infty \\ &\leq A_* \|s_{CS}\|_\infty \qquad \text{by (44) and (7)} \\ &= A_* \left\|\sum_{I\in\Lambda_M} \beta_{n,I}\varphi_I\right\|_\infty \\ &\leq A_* r_M\sqrt{D_M} \max_{I\in\Lambda_M} |\beta_{n,I}| \qquad \text{by (92)} \\ &\leq \tilde{R}_{n,D_M,\alpha} \qquad \text{by (126)} \end{aligned} \tag{130}$$

Now, by observing that

$$\begin{aligned} \left\|\psi_{1,M}\cdot((s_M \times s_{CS} + s_M) - s_M)\right\|_2^2 &= \|s_{CS}\|_2^2 \\ &= 2rC \ , \end{aligned}$$

we get by (125) and (130) that for all $n \geq n_0(B_2, A_-, r_M, \beta)$, $(s_M \times s_{CS} + s_M) \in \mathcal{F}_{(C,rC]}$ with probability at least $1 - (2D_M+1)n^{-\beta}$. ■

We are now ready to prove the lower bound (124) for the expected value of the largest increment of the empirical process over $\mathcal{F}_{(C,rC]}$.

Proof of Lemma 19. Let us begin with the lower bound of $\mathbb{E}^{\frac{1}{2}}\left(\sup_{s\in\mathcal{F}_{(C,rC]}}(P_n-P)\left(\psi_{1,M}\cdot(s_M-s)\right)\right)^2$, a result that will be needed further in the proof. By Lemma 20, if we set

$$\tilde{\Omega}=\left\{(s_M\times s_{CS}+s_M)\in\mathcal{F}_{(C,rC]}\right\}$$

if we choose $\beta=4$ and if

$$A_\infty\geq 64B_2\sqrt{A_u}A_*r_M\ ,$$

then it holds, for all $n\geq n_0\left(B_2,A_+,A_-,r_M\right)$,

$$\mathbb{P}\left[\tilde{\Omega}\right]\geq 1-\frac{2D_M+1}{n^4}\ . \tag{131}$$

Also, it holds

$$\begin{aligned}
&\mathbb{E}^{\frac{1}{2}}\left(\sup_{s\in\mathcal{F}_{(C,rC]}}(P_n-P)\left(\psi_{1,M}\cdot(s_M-s)\right)\right)^2\\
&\geq\mathbb{E}^{\frac{1}{2}}\left[\left((P_n-P)\left(\sum_{I\in\Lambda_M}\beta_{n,I}\varphi_I\right)\right)^2\mathbf{1}_{\tilde{\Omega}}\right]\\
&\geq\sqrt{2rC}\sqrt{\mathbb{E}\left[\left(\sum_{I\in\Lambda_M}(P_n-P)^2(\varphi_I)\right)\mathbf{1}_{\tilde{\Omega}}\right]}\ .
\end{aligned}\tag{132}$$

Furthermore, since by (92) $\|\varphi_I\|_\infty\leq\sqrt{D_M}r_M$ for all $I\in\Lambda_M$, and since $P(\varphi_I)\geq 0$ we have

$$\left|\sum_{I\in\Lambda_M}(P_n-P)^2(\varphi_I)\right|\leq D_M\max_{I\in\Lambda_M}\|\varphi_I\|_\infty^2\leq r_M^2D_M^2$$

and it ensures by (131), for all $n\geq n_0\left(B_2,A_+,A_-,r_M\right)$,

$$\mathbb{E}\left[\left(\sum_{I\in\Lambda_M}(P_n-P)^2(\varphi_I)\right)\mathbf{1}_{\tilde{\Omega}}\right]\geq\mathbb{E}\left[\left(\sum_{I\in\Lambda_M}(P_n-P)^2(\varphi_I)\right)\right]-r_M^2D_M^2\frac{2D_M+1}{n^4}\ .$$

Comparing the last inequality with (132), we obtain the lower bound, for all $n\geq n_0\left(B_2,A_+,A_-,r_M\right)$,

$$\begin{aligned}
&\mathbb{E}^{\frac{1}{2}}\left(\sup_{s\in\mathcal{F}_{(C,rC]}}(P_n-P)\left(\psi_{1,M}\cdot(s_M-s)\right)\right)^2\\
&\geq\sqrt{2rC}\sqrt{\mathbb{E}\left[\sum_{I\in\Lambda_M}(P_n-P)^2(\varphi_I)\right]}-r_MD_M\sqrt{2rC}\sqrt{\frac{2D_M+1}{n^4}}\\
&=\sqrt{\frac{2rC\left(D_M-1\right)}{n}}-r_MD_M\sqrt{2rC}\sqrt{\frac{2D_M+1}{n^4}}\ .
\end{aligned}$$

Now, since $D_M\leq A_+n\left(\ln n\right)^2$, we get for all $n\geq n_0\left(A_+,r_M\right)$,

$$r_MD_M\sqrt{2rC}\sqrt{\frac{2D_M+1}{n^4}}\leq\frac{1}{\sqrt{D_M}}\times\sqrt{\frac{2rC\left(D_M-1\right)}{n}}$$

and so, if $A_\infty\geq 64B_2\sqrt{A_u}A_*r_M$ then, for all $n\geq n_0\left(B_2,A_+,A_-,r_M\right)$,

$$\mathbb{E}^{\frac{1}{2}}\left(\sup_{s\in\mathcal{F}_{(C,rC]}}(P_n-P)\left(\psi_{1,M}\cdot(s_M-s)\right)\right)^2\geq\left(1-\frac{1}{\sqrt{D_M}}\right)\sqrt{\frac{2rC\left(D_M-1\right)}{n}}\ . \tag{133}$$

Now, as $D_M \geq A_- (\ln n)^2$ we have for all $n \geq n_0(A_-)$, $D_M^{-1/2} \leq 1/2$. Moreover we have $rC \geq A_l D_M n^{-1}$, so we deduce from (133) that, for all $n \geq n_0(B_2, A_+, A_-, r_M)$,

$$\mathbb{E}^{\frac{1}{2}} \left(\sup_{s \in \mathcal{F}_{(C,rC]}} (P_n - P) \left(\psi_{1,M} \cdot (s_M - s) \right) \right)^2 \geq \sqrt{\frac{A_l}{2}} \frac{D_M}{n} \ . \tag{134}$$

We turn now to the lower bound of $\mathbb{E}\left[\sup_{s \in \mathcal{F}_{(C,rC]}} (P_n - P)\left(\psi_{1,M} \cdot (s_M - s)\right)\right]$. First observe that $s \in \mathcal{F}_{(C,rC]}$ implies that $2s_M - s \in \mathcal{F}_{(C,rC]}$, so that

$$\mathbb{E}\left[\sup_{s \in \mathcal{F}_{(C,rC]}} (P_n - P) \left(\psi_{1,M} \cdot (s_M - s) \right) \right] = \mathbb{E}\left[\sup_{s \in \mathcal{F}_{(C,rC]}} \left| (P_n - P) \left(\psi_{1,M} \cdot (s_M - s) \right) \right| \right] \ . \tag{135}$$

In the next step, we apply Corollary 32. More precisely, using notations of Corollary 32, we set

$$\mathcal{F} = \left\{ \psi_{1,M} \cdot (s_M - s) \, , \ s \in \mathcal{F}_{(C,rC]} \right\}$$

and

$$Z = \sup_{s \in \mathcal{F}_{(C,rC]}} \left| (P_n - P) \left(\psi_{1,M} \cdot (s_M - s) \right) \right| \ .$$

Now, since for all $n \geq n_0(A_+, A_\infty)$, it holds $\tilde{R}_{n,D_M,\alpha} \leq 1/2$, we get by (45),

$$\sup_{f \in \mathcal{F}} \|f - Pf\|_\infty \leq 2 \sup_{s \in \mathcal{F}_{(C,rC]}} \left\| \psi_{1,M} \cdot (s_M - s) \right\|_\infty \leq 2 A_{\min}^{-1} \tilde{R}_{n,D_M,\alpha} \leq A_{\min}^{-1} \ .$$

we set $b = A_{\min}^{-1}$. Since we assume that $rC \leq A_u \frac{D_M}{n}$, it moreover holds

$$\sup_{f \in \mathcal{F}} \mathrm{Var}(f) \leq \sup_{s \in \mathcal{F}_{(C,rC]}} P\left(\psi_{1,M} \cdot (s_M - s) \right)^2 \leq 2rC \leq 2A_u \frac{D_M}{n}$$

and so we set $\sigma^2 = 2A_u \frac{D_M}{n}$. Now, by (134) we have, for all $n \geq n_0(B_2, A_+, A_-, r_M)$,

$$\sqrt{\mathbb{E}\left[Z^2\right]} \geq \sqrt{\frac{A_l}{2}} \frac{D_M}{n} \ . \tag{136}$$

Hence, a positive constant $L_{A_l, A_u, A_{\min}}$ exists such that, by setting

$$\varkappa_n = \frac{L_{A_l, A_u, A_{\min}}}{\sqrt{D_M}}$$

we can get using (136), that for all $n \geq n_0(B_2, A_+, A_-, r_M)$,

$$\varkappa_n^2 \mathbb{E}\left[Z^2\right] \geq \frac{\sigma^2}{n}$$

$$\varkappa_n^2 \sqrt{\mathbb{E}\left[Z^2\right]} \geq \frac{b}{n}$$

and that, as $D_M \geq A_- (\ln n)^2$, we have for all $n \geq n_0(A_l, A_u, A_-, A_{\min})$,

$$\varkappa_n \in (0,1) \ .$$

So, using (135) and Corollary 32, it holds for all $n \geq n_0(B_2, A_+, A_-, A_l, A_u, r_M, A_{\min})$,

$$\mathbb{E}\left[\sup_{s \in \mathcal{F}_{(C,rC]}} (P_n - P) \left(\psi_{1,M} \cdot (s_M - s) \right) \right] \geq \left(1 - \frac{L_{A_l, A_u, A_{\min}}}{\sqrt{D_M}} \right) \mathbb{E}^{\frac{1}{2}} \left(\sup_{s \in \mathcal{F}_{(C,rC]}} (P_n - P) \left(\psi_{1,M} \cdot (s_M - s) \right) \right)^2 . \tag{137}$$

Finally, using (133) in the right-hand side of Inequality (137), we can deduce that for all $n \geq n_0 (B_2, A_+, A_-, A_l, A_u, r_M, A_{\min})$,

$$\mathbb{E}\left[\sup_{s \in \mathcal{F}_{(C,rC]}} (P_n - P)\left(\psi_{1,M} \cdot (s_M - s)\right)\right] \geq \left(1 - \frac{L_{A_l,A_u,A_{\min}}}{\sqrt{D_M}}\right)\sqrt{\frac{2rC\left(D_M - 1\right)}{n}}$$

and so (124) is proved. ■
The two following lemmas give some controls of the supremum over the second order terms in the expansion of the contrast (81).

Lemma 21 *Let $C \geq 0$ and $A_+ > 0$. Under (45), assuming that*

$$A_+ \frac{n}{(\ln n)^2} \geq D_M \ ,$$

it holds, for all $n \geq n_0 (A_+, A_{\min}, A_\infty)$,

$$\mathbb{E}\left[\sup_{s \in \mathcal{F}_C} \left|(P_n - P)\left(\psi_2 \circ \left(\frac{s - s_M}{s_M}\right)\right)\right|\right] \leq 8A_{\min}^{-2}\tilde{R}_{n,D_M,\alpha}\sqrt{\frac{2C\left(D_M - 1\right)}{n}} \ .$$

Proof. We define the Rademacher process $\mathcal{R}_n$ on a class $\mathcal{F}$ of measurable functions from $\mathcal{Z}$ to $\mathbb{R}$, to be

$$\mathcal{R}_n (f) = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(\xi_i) \ , \quad f \in \mathcal{F}$$

where ε_i are independent Rademacher random variables also independent from the ξ_i. By the usual symmetrization argument we have

$$\mathbb{E}\left[\sup_{s \in \mathcal{F}_C} \left|(P_n - P)\left(\psi_2 \circ \left(\frac{s - s_M}{s_M}\right)\right)\right|\right] \leq 2\mathbb{E}\left[\sup_{s \in \mathcal{F}_C} \left|\mathcal{R}_n\left(\psi_2 \circ \left(\frac{s - s_M}{s_M}\right)\right)\right|\right] \ . \tag{138}$$

As $A_+ \frac{n}{(\ln n)^2} \geq D_M$, we have, for all $n \geq n_0 (A_+, A_{\min}, A_\infty)$,

$$\tilde{R}_{n,D_M,\alpha} \leq \frac{A_{\min}}{2} \ .$$

Hence, by Inequality (19) of Lemma 6 it holds for all $n \geq n_0 (A_+, A_{\min}, A_\infty)$, for all $(x, y) \in \left[-A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha},\ A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha}\right]^2$,

$$|\psi_2 (x) - \psi_2 (y)| \leq 2A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha}\,|x - y| \ . \tag{139}$$

We define now the following real-valued function ρ,

$$\rho (x) = \begin{cases} \left(2A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha}\right)^{-1}\psi_2 (x) & \text{if } x \in \left[-A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha},\ A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha}\right] \\ \left(2A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha}\right)^{-1}\psi_2\left(-A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha}\right) & \text{if } x \leq -A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha} \\ \left(2A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha}\right)^{-1}\psi_2\left(A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha}\right) & \text{if } x \geq A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha} \end{cases}$$

and since $\rho (0) = h (0) = 0$, it follows from (139) that ρ is a contraction mapping for all $n \geq n_0 (A_+, A_{\min}, A_\infty)$. Then, taking the expectation with respect to the Rademacher variables, we then get for all $n \geq n_0 (A_+, A_{\min}, A_\infty)$,

$$\begin{aligned} &\mathbb{E}_\varepsilon\left[\sup_{s \in \mathcal{F}_C} \left|\mathcal{R}_n\left(\psi_2 \circ \left(\frac{s - s_M}{s_M}\right)\right)\right|\right] \\ &= 2A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha}\mathbb{E}_\varepsilon\left[\sup_{s \in \mathcal{F}_C} \left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \rho\left(\left(\frac{s - s_M}{s_M}\right)(\xi_i)\right)\right|\right] \end{aligned} \tag{140}$$

We can now apply Theorem 28 to get for all $n \geq n_0\left(A_+, A_{\min}, A_\infty\right)$,

$$\mathbb{E}_\varepsilon\left[\sup_{s\in\mathcal{F}_C}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \rho\left(\left(\frac{s-s_M}{s_M}\right)(\xi_i)\right)\right|\right] \leq 2\mathbb{E}_\varepsilon\left[\sup_{s\in\mathcal{F}_C}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \left(\frac{s-s_M}{s_M}\right)(\xi_i)\right|\right]$$
$$= 2\mathbb{E}_\varepsilon\left[\sup_{s\in\mathcal{F}_C}\left|\mathcal{R}_n\left(\frac{s-s_M}{s_M}\right)\right|\right] \tag{141}$$

and so we derive successively the following upper bounds in mean, for all $n \geq n_0\left(A_+, A_{\min}, A_\infty\right)$,

$$\mathbb{E}\left[\sup_{s\in\mathcal{F}_C}\left|\mathcal{R}_n\left(\psi_2\circ\left(\frac{s-s_M}{s_M}\right)\right)\right|\right] = \mathbb{E}\left[\mathbb{E}_\varepsilon\left[\sup_{s\in\mathcal{F}_C}\left|\mathcal{R}_n\left(\psi_2\circ\left(\frac{s-s_M}{s_M}\right)\right)\right|\right]\right]$$
$$\leq 2A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha}\mathbb{E}\left[\mathbb{E}_\varepsilon\left[\sup_{s\in\mathcal{F}_C}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\rho\left(\left(\frac{s-s_M}{s_M}\right)(\xi_i)\right)\right|\right]\right] \qquad \text{by (140)}$$
$$\leq 4A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha}\mathbb{E}\left[\sup_{s\in\mathcal{F}_C}\left|\mathcal{R}_n\left(\frac{s-s_M}{s_M}\right)\right|\right] \qquad \text{by (141)}$$
$$\leq 4A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha}\sqrt{\mathbb{E}\left[\left(\sup_{s\in\mathcal{F}_C}\left|\mathcal{R}_n\left(\frac{s-s_M}{s_M}\right)\right|\right)^2\right]}\,. \tag{142}$$

Hence, for all $n \geq n_0\left(A_+, A_{\min}, A_\infty\right)$,

$$\sqrt{\mathbb{E}\left[\left(\sup_{s\in\mathcal{F}_C}\left|\mathcal{R}_n\left(\frac{s-s_M}{s_M}\right)\right|\right)^2\right]}$$
$$= \sqrt{\mathbb{E}\left[\left(\sup_{s\in\mathcal{F}_C}\left|\mathcal{R}_n\left(\psi_{1,M}\cdot(s_M-s)\right)\right|\right)^2\right]}$$
$$\leq \sqrt{\mathbb{E}\left[\left(\sup\left\{\left|\sum_{I\in\Lambda_M} a_I\mathcal{R}_n\left(\varphi_I\right)\right|;\ \sum_{I\in\Lambda_M} a_I^2 \leq 2C\right\}\right)^2\right]}$$
$$= \sqrt{2C}\sqrt{\mathbb{E}\left[\sum_{I\in\Lambda_M}\left(\mathcal{R}_n\left(\varphi_I\right)\right)^2\right]} = \sqrt{\frac{2C\left(D_M-1\right)}{n}} \tag{143}$$

and the result follows by injecting (142) and (143) in (138). ■

Lemma 22 *Let $A_+, A_-, A_l, \beta, C_- > 0$, and assume (45) and (**Alr**). Then if $C_- \geq A_l\frac{D_M}{n}$ and $A_+ n\left(\ln n\right)^{-2} \geq D_M \geq A_-\left(\ln n\right)^2$, then a positive constant $L_{A_-,A_l,\beta}$ exists such that, for all $n \geq n_0\left(A_{\min}, A_\infty, A_+, A_l\right)$,*

$$\mathbb{P}\left[\forall C > C_-,\quad \sup_{s\in\mathcal{F}_C}\left|\left(P_n - P\right)\left(\psi_2\circ\left(\frac{s-s_M}{s_M}\right)\right)\right| \leq L_{A_-,A_l,A_{\min},\beta}\sqrt{\frac{C\left(D_M-1\right)}{n}}\tilde{R}_{n,D_M,\alpha}\right] \geq 1 - n^{-\beta}\ .$$

Proof. First notice that, as $A_+ n\left(\ln n\right)^{-2} \geq D_M$,

$$\tilde{R}_{n,D_M,\alpha} \leq \frac{A_\infty\sqrt{A_+}}{\sqrt{\ln n}}\ .$$

As a consequence, for all $n \geq n_0\left(A_{\min}, A_\infty, A_+\right)$,

$$\tilde{R}_{n,D_M,\alpha} \leq \sqrt{2}A_{\min}\ . \tag{144}$$

Now, since $\cup_{C>C_-}\mathcal{F}_C \subset B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D_M,\alpha}\right)$ where

$$B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D_M,\alpha}\right) = \left\{s \in M, \|s - s_M\|_\infty \leq \tilde{R}_{n,D_M,\alpha}\right\},$$

we have by (144) and (45), for all $s \in \cup_{C>C_-}\mathcal{F}_C$ and for all $n \geq n_0\left(A_{\min}, A_\infty, A_+\right)$,

$$\begin{aligned}
&\frac{1}{2}\left\|\psi_{1,M}\cdot(s - s_M)\right\|_2^2 \\
&\leq \frac{A_{\min}^{-2}}{2}\|s - s_M\|_\infty^2 \\
&\leq \frac{A_{\min}^{-2}}{2}\tilde{R}_{n,D_M,\alpha}^2 \leq 1 \ .
\end{aligned}$$

We thus have, for all $n \geq n_0\left(A_{\min}, A_\infty, A_+\right)$,

$$\cup_{C>C_-}\mathcal{F}_C = \cup_{C_-\wedge 1<C\leq 1}\mathcal{F}_C$$

and by monotonicity of the collection $\mathcal{F}_C$, for some $q > 1$ and $J = \left\lfloor \frac{|\ln(C_-\wedge 1)|}{\ln q}\right\rfloor + 1$, it holds

$$\cup_{C_-\wedge 1<C\leq 1}\mathcal{F}_C \subset \cup_{j=0}^{J}\mathcal{F}_{q^jC_-} \ .$$

Simple computations show that, since $D_M \geq 1$ and $C_- \geq A_l\frac{D_M}{n} \geq \frac{A_l}{n}$, one can find a constant $L_{A_l,q}$ such that

$$J \leq L_{A_l,q}\ln n \ .$$

Moreover, by monotonicity of $C \longmapsto \sup_{s\in\mathcal{F}_C}\left|(P_n - P)\left(\psi_2 \circ \left(\frac{s-s_M}{s_M}\right)\right)\right|$, we have uniformly in $C \in \left(q^{j-1}C_-, q^jC_-\right]$,

$$\sup_{s\in\mathcal{F}_C}\left|(P_n - P)\left(\psi_2 \circ \left(\frac{s - s_M}{s_M}\right)\right)\right| \leq \sup_{s\in\mathcal{F}_{q^jC_-}}\left|(P_n - P)\left(\psi_2 \circ \left(\frac{s - s_M}{s_M}\right)\right)\right| \ .$$

Hence we get, for all $n \geq n_0\left(A_{\min}, A_\infty, A_+\right)$ and any $L > 0$,

$$\begin{aligned}
&\mathbb{P}\left[\forall C > C_-, \quad \sup_{s\in\mathcal{F}_C}\left|(P_n - P)\left(\psi_2 \circ \left(\frac{s - s_M}{s_M}\right)\right)\right| \leq L\sqrt{\frac{2C\left(D_M - 1\right)}{n}}\tilde{R}_{n,D_M,\alpha}\right] \\
&\geq \mathbb{P}\left[\forall j \in \{1, ..., J\}, \quad \sup_{s\in\mathcal{F}_{q^jC_-}}\left|(P_n - P)\left(\psi_2 \circ \left(\frac{s - s_M}{s_M}\right)\right)\right| \leq L\sqrt{\frac{2q^jC_-\left(D_M - 1\right)}{n}}\tilde{R}_{n,D_M,\alpha}\right] \ .
\end{aligned}$$

Now, for all $n \geq n_0\left(A_{\min}, A_\infty, A_+\right)$ and any $L > 0$,

$$\begin{aligned}
&\mathbb{P}\left[\forall j \in \{1, ..., J\}, \quad \sup_{s\in\mathcal{F}_{q^jC_-}}\left|(P_n - P)\left(\psi_2 \circ \left(\frac{s - s_M}{s_M}\right)\right)\right| \leq L\sqrt{\frac{2q^jC_-\left(D_M - 1\right)}{n}}\tilde{R}_{n,D_M,\alpha}\right] \\
&= 1 - \mathbb{P}\left[\exists j \in \{1, ..., J\}, \quad \sup_{s\in\mathcal{F}_{q^jC_-}}\left|(P_n - P)\left(\psi_2 \circ \left(\frac{s - s_M}{s_M}\right)\right)\right| > L\sqrt{\frac{2q^jC_-\left(D_M - 1\right)}{n}}\tilde{R}_{n,D_M,\alpha}\right] \\
&\geq 1 - \sum_{j=1}^{J}\mathbb{P}\left[\sup_{s\in\mathcal{F}_{q^jC_-}}\left|(P_n - P)\left(\psi_2 \circ \left(\frac{s - s_M}{s_M}\right)\right)\right| > L\sqrt{\frac{2q^jC_-\left(D_M - 1\right)}{n}}\tilde{R}_{n,D_M,\alpha}\right] . \qquad (145)
\end{aligned}$$

Given $j \in \{1, ..., J\}$, Lemma 21 yields

$$\mathbb{E}\left[\sup_{s\in\mathcal{F}_{q^jC_-}}\left|(P_n - P)\left(\psi_2 \circ \left(\frac{s - s_M}{s_M}\right)\right)\right|\right] \leq 8A_{\min}^{-2}\sqrt{\frac{2q^jC_-\left(D_M - 1\right)}{n}}\tilde{R}_{n,D_M,\alpha} \ ,$$

and we can next apply Bousquet's inequality (173) to handle the deviations around the mean. Since for all $n \geq n_0 (A_{\min}, A_\infty, A_+)$ we have for all $s \in \mathcal{F}_{q^j C_-}$,

$$\|s - s_M\|_\infty \leq \tilde{R}_{n,D_M,\alpha} \leq \frac{A_{\min}}{2}$$

we can apply Inequalities (45) and (83) to get, for all $n \geq n_0 (A_{\min}, A_\infty, A_+)$,

$$\begin{aligned}
\sup_{s \in \mathcal{F}_{q^j C_-}} \left\| \psi_2 \circ \left(\frac{s - s_M}{s_M} \right) - P \left(\psi_2 \circ \left(\frac{s - s_M}{s_M} \right) \right) \right\| &\leq 2 \sup_{s \in \mathcal{F}_{q^j C_-}} \left\| \psi_2^s \cdot (s - s_M) \right\|_\infty \\
&\leq 2 A_{\min}^{-1} \sup_{s \in \mathcal{F}_{q^j C_-}} \left\| \frac{1}{s_M} (s - s_M)^2 \right\|_\infty \\
&\leq 2 A_{\min}^{-2} \tilde{R}_{n,D_M,\alpha}^2
\end{aligned}$$

and, for all $s \in \mathcal{F}_{q^j C_-}$, for all $n \geq n_0 (A_{\min}, A_\infty, A_+)$,

$$\begin{aligned}
&\mathrm{Var} \left(\psi_2 \circ \left(\frac{s - s_M}{s_M} \right) \right) \\
&\leq P \left[\left(\psi_2 \circ \left(\frac{s - s_M}{s_M} \right) \right)^2 \right] \\
&\leq A_{\min}^{-2} \|s - s_M\|_\infty^2 P \left[\left(\frac{s - s_M}{s_M} \right)^2 \right] \qquad \text{by (83)} \\
&\leq 2 A_{\min}^{-2} \tilde{R}_{n,D_M,\alpha}^2 q^j C_- \ .
\end{aligned}$$

It follows that Inequality (173) applied with $\varepsilon = 1$ gives, for all $x > 0$ and for all $n \geq n_0 (A_{\min}, A_\infty, A_+)$,

$$\mathbb{P} \left[\begin{array}{c} \sup_{s \in \mathcal{F}_{q^j C_-}} \left| (P_n - P) \left(\psi_2 \circ \left(\frac{s - s_M}{s_M} \right) \right) \right| \geq \\ 16 A_{\min}^{-2} \sqrt{\frac{2 q^j C_- (D_M - 1)}{n}} \tilde{R}_{n,D_M,\alpha} + \sqrt{\frac{4 A_{\min}^{-2} \tilde{R}_{n,D_M,\alpha}^2 q^j C_- x}{n}} + \frac{8 A_{\min}^{-2} \tilde{R}_{n,D_M,\alpha}^2 x}{3n} \end{array} \right] \leq \exp(-x) \ . \tag{146}$$

As a consequence, as $D_M \geq A_- (\ln n)^2$, $C_- \geq A_l D_M n^{-1}$ and as $\tilde{R}_{n,D_M,\alpha} \leq 1$ for all $n \geq n_0 (A_{\min}, A_\infty, A_+)$, taking $x = \gamma \ln n$ in (146) for some $\gamma > 0$, easy computations show that a positive constant $L_{A_-, A_l, A_{\min}, \gamma}$ independent of j exists such that for all $n \geq n_0 (A_{\min}, A_\infty, A_+)$,

$$\mathbb{P} \left[\sup_{s \in \mathcal{F}_{q^j C_-}} \left| (P_n - P) \left(\psi_2 \circ \left(\frac{s - s_M}{s_M} \right) \right) \right| \geq L_{A_-, A_l, A_{\min}, \gamma} \sqrt{\frac{q^j C_- (D_M - 1)}{n}} \tilde{R}_{n,D_M,\alpha} \right] \leq \frac{1}{n^\gamma} \ .$$

Hence, using (145), we get for all $n \geq n_0 (A_{\min}, A_\infty, A_+)$,

$$\mathbb{P} \left[\forall C > C_-, \quad \sup_{s \in \mathcal{F}_C} \left| (P_n - P) \left(\psi_2 \circ \left(\frac{s - s_M}{s_M} \right) \right) \right| \leq L_{A_-, A_l, A_{\min}, \gamma} \sqrt{\frac{2C (D_M - 1)}{n}} \tilde{R}_{n,D_M,\alpha} \right] \geq 1 - \frac{J}{n^\gamma} \ .$$

And finally, as $J \leq L_{A_l, q} \ln n$, taking $\gamma = \beta + 1$ and $q = 2$ gives the result for all $n \geq n_0 (A_{\min}, A_\infty, A_+, A_l)$. ■

Having controlled the residual empirical process driven by the remainder terms in the contrast, and having proved sharp bounds for the expectation of the increments of the main empirical process on our slices, it remains to combine the above lemmas in order to establish the crucial probability estimates controlling the empirical excess risk on the slides.

Lemma 23 *Let $\beta, A_-, A_+, A_l, C > 0$. Assume that (45) and (**Alr**) hold. A positive constant A_4 exists, only depending on $A_{\min}, A_\infty, r_M, \beta$, such that, if*

$$A_l \frac{D_M}{n} \leq C \leq (1 + A_4 \nu_n)^2 \frac{D_M - 1}{2n} \quad \text{and} \quad A_+ \frac{n}{(\ln n)^2} \geq D_M \geq A_- (\ln n)^2$$

where $\nu_n = \max\left\{\sqrt{\frac{\ln n}{D_M}},\ \sqrt{\frac{D_M \ln n}{n}}\right\}$, then for all $n \geq n_0 (A_l, A_-, A_+, A_{\min}, r_M, A_\infty, \beta)$,

$$\mathbb{P}\left[\sup_{s \in \mathcal{F}_C} P_n (K s_M - K s) \geq \left(1 + L_{A_-, A_l, A_{\min}, A_\infty, r_M, \beta} \times \nu_n\right) \sqrt{\frac{2C (D_M - 1)}{n}} - C\right] \leq 2 n^{-\beta} \ .$$

Proof. Start with

$$\begin{aligned}
&\sup_{s \in \mathcal{F}_C} P_n (K s_M - K s) \\
&= \sup_{s \in \mathcal{F}_C} \left\{ P_n \left(\psi_{1,M} \cdot (s_M - s) - \psi_2 \circ \left(\frac{s - s_M}{s_M} \right) \right) \right\} \\
&= \sup_{s \in \mathcal{F}_C} \left\{ (P_n - P) \left(\psi_{1,M} \cdot (s_M - s) \right) - (P_n - P) \left(\psi_2 \circ \left(\frac{s - s_M}{s_M} \right) \right) - P (K s - K s_M) \right\} \\
&\leq \sup_{s \in \mathcal{F}_C} \left\{ (P_n - P) \left(\psi_{1,M} \cdot (s_M - s) \right) - P (K s - K s_M) \right\} \\
&+ \sup_{s \in \mathcal{F}_C} \left| (P_n - P) \left(\psi_2 \circ \left(\frac{s - s_M}{s_M} \right) \right) \right| \ . \qquad (147)
\end{aligned}$$

Recall that by (89) we have, for all $s \in \mathcal{F}_C$ and for all $n \geq n_0 (A_+, A_{\min}, A_\infty)$,

$$P (K s - K s_M) \geq \left(1 - \frac{4}{3 A_{\min}} \tilde{R}_{n, D_M, \alpha} \right) \frac{1}{2} \left\| \psi_{1,M} (s - s_M) \right\|_2^2 \ . \qquad (148)$$

Next, recall that

$$D_L = \left\{ s \in \widetilde{M} \ ; \ \frac{1}{2} \left\| \psi_{1,M} \cdot (s - s_M) \right\|_2^2 = L \right\} \bigcap B_{(M, L_\infty)} \left(s_M, \tilde{R}_{n, D_M, \alpha} \right) \ .$$

Moreover, we notice that, for any $s \in M$,

$$\psi_{1,M} (s - s_M) = \frac{s - s_M}{s_M}$$

is a piecewise constant function with respect to the partition Λ_M. Thus $\psi_{1,M} \cdot (s - s_M) \in M$ for any $s \in M$, and we have

$$\begin{aligned}
&\sup_{s \in D_L} (P_n - P) \left(\psi_{1,M} \cdot (s_M - s) \right) \\
&\leq \sup_{\left\{ s \in M, \|t\|_2^2 = 2L \right\}} (P_n - P) (t) \\
&\leq \sqrt{2L} \sqrt{\sum_{I \in \Lambda_M} (P_n - P)^2 (\varphi_I)}
\end{aligned}$$

where the last bound follows from Cauchy-Schwarz inequality. Then, for all $n \geq n_0\left(A_+, A_{\min}, A_\infty\right)$,

$$\begin{aligned}
&\sup_{s\in\mathcal{F}_C}\left\{\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(s_M-s\right)\right)-P\left(Ks-Ks_M\right)\right\}\\
&\leq \sup_{L\leq C}\sup_{s\in D_L}\left\{\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(s_M-s\right)\right)-\left(1-\frac{4}{3A_{\min}}\tilde{R}_{n,D_M,\alpha}\right)L\right\} \qquad \text{by (148)}\\
&\leq \sup_{L\leq C}\left\{\sqrt{2L}\sqrt{\sum_{I\in\Lambda_M}\left(P_n-P\right)^2\left(\varphi_I\right)}-\left(1-\frac{4}{3A_{\min}}\tilde{R}_{n,D_M,\alpha}\right)L\right\}\ .
\end{aligned}$$

Hence, since $D_M \geq A_-\left(\ln n\right)^2 \geq 2$ for all $n \geq n_0\left(A_-\right)$, we deduce from Lemma 18 that for all $n \geq n_0\left(A_-, A_+, A_{\min}, A_\infty\right)$,

$$\mathbb{P}\left[\begin{array}{c}\sup_{s\in\mathcal{F}_C}\left\{\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(s_M-s\right)\right)-P\left(Ks-Ks_M\right)\right\}\\ \geq \sup_{L\leq C}\left\{\sqrt{2L}\left(1+\tau_n\right)\sqrt{\frac{D_M-1}{n}}-\left(1-\frac{4}{3A_{\min}}\tilde{R}_{n,D_M,\alpha}\right)L\right\}\end{array}\right] \leq n^{-\beta}\ . \tag{149}$$

where

$$\begin{aligned}
\tau_n &= L_{r_M,\beta}\left(\sqrt{\frac{\ln n}{D_M}}\vee\frac{\sqrt{\ln n}}{n^{1/4}}\right)\\
&\leq L_{r_M,\beta}\left(\sqrt{\frac{\ln n}{D_M}}\vee\sqrt{\frac{D_M\ln n}{n}}\right)\\
&\leq L_{r_M,\beta}\nu_n\ .
\end{aligned} \tag{150}$$

Assume now that

$$C \leq \frac{D_M-1}{n}\ . \tag{151}$$

then we have for all $0 \leq L \leq C$,

$$\frac{4}{3A_{\min}}\tilde{R}_{n,D_M,\alpha}\times L \leq L_{A_{\min},A_\infty}\sqrt{\frac{D_M\ln n}{n}}\times\sqrt{L}\sqrt{\frac{D_M-1}{n}} \leq L_{A_{\min},A_\infty}\nu_n\sqrt{L}\sqrt{\frac{D_M-1}{n}}\ . \tag{152}$$

Hence, using (150) and (152) in (149), if $C \leq \frac{D_M-1}{n}$ it holds for all $n \geq n_0\left(A_-, A_+, A_{\min}, A_\infty\right)$,

$$\mathbb{P}\left[\begin{array}{c}\sup_{s\in\mathcal{F}_C}\left\{\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(s_M-s\right)\right)-P\left(Ks-Ks_M\right)\right\}\\ \geq \sup_{L\leq C}\left\{\sqrt{2L}\left(1+L_{A_{\min},A_\infty,r_M,\beta}\nu_n\right)\sqrt{\frac{D_M-1}{n}}-L\right\}\end{array}\right] \leq n^{-\beta}\ . \tag{153}$$

Now, we set $A_4 = L_{A_{\min},A_\infty,r_M,\beta}$ the positive constant appearing in (153). If $C \leq \left(1+A_4\nu_n\right)^2\frac{D_M-1}{2n}$ then for all $n \geq n_0\left(A_4\right)$ (151) is satisfied and we get after simple calculations that

$$\sup_{L\leq C}\left\{\sqrt{2L}\left(1+A_4\nu_n\right)\sqrt{\frac{D_M-1}{n}}-L\right\} = \sqrt{2C}\left(1+A_4\nu_n\right)\sqrt{\frac{D_M-1}{n}}-C$$

and as a consequence, for all $n \geq n_0\left(A_-, A_+, A_{\min}, A_4, A_\infty\right)$,

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_C}\left\{\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(s_M-s\right)\right)-P\left(Ks-Ks_M\right)\right\} \geq \sqrt{2C}\left(1+A_4\nu_n\right)\sqrt{\frac{D_M-1}{n}}-C\right] \leq n^{-\beta}\ . \tag{154}$$

Moreover, since $C \geq A_l\frac{D_M}{n}$, we can derive from Lemma 22 that for all $n \geq n_0\left(A_{\min}, A_\infty, A_+, A_l\right)$,

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_C}\left|\left(P_n-P\right)\left(\psi_2\circ\left(\frac{s-s_M}{s_M}\right)\right)\right| \geq L_{A_-,A_l,A_{\min},\gamma}\sqrt{\frac{C\left(D_M-1\right)}{n}}\tilde{R}_{n,D_M,\alpha}\right] \leq n^{-\beta}$$

and as

$$\tilde{R}_{n,D_M,\alpha} \leq L_{A_\infty} \nu_n$$

we have, for all $n \geq n_0 (A_{\min}, A_\infty, A_+, A_l)$,

$$\mathbb{P}\left[\sup_{s \in \mathcal{F}_C} \left| (P_n - P) \left(\psi_2 \circ \left(\frac{s - s_M}{s_M} \right) \right) \right| \geq L_{A_-, A_l, A_{\min}, A_\infty} \sqrt{\frac{C (D_M - 1)}{n}} \tilde{R}_{n,D_M,\alpha} \right] \leq n^{-\beta} \ . \tag{155}$$

The conclusion follows by making use of (154) and (155) in Inequality (147). ■

Lemma 24 *Let $\beta, A_-, A_+, A_u, C \geq 0$. Assume that (45) and (**Alr**) hold. A positive constant A_5, depending on $A_\infty, r_M, A_{\min}, A_-, A_u, \beta$, exists such that, if it holds*

$$A_u \frac{D_M}{n} \geq C \geq \frac{1}{4} (1 + A_5 \nu_n)^2 \frac{D_M - 1}{n} \quad \text{and} \quad A_+ \frac{n}{(\ln n)^2} \geq D_M \geq A_- (\ln n)^2$$

where $\nu_n = \max \left\{ \sqrt{\frac{\ln n}{D_M}}, \sqrt{\frac{D_M \ln n}{n}} \right\}$, then for all $n \geq n_0 (A_\infty, A_{cons}, n_1, A_+, \alpha)$,

$$\mathbb{P}\left[\sup_{s \in \mathcal{F}_{>C}} P_n (K s_M - K s) \geq (1 + A_5 \nu_n) \sqrt{\frac{2C (D_M - 1)}{n}} - C \right] \leq 2n^{-\beta} \ .$$

Moreover, when we only assume $C \geq 0$ (and keep the other assumptions unchanged), a positive constant A_6 exists, depending only on $A_\infty, r_M, A_{\min}, A_-, \beta$, such that we have for all $n \geq n_0 \left(A_{\min}, A_\infty, A_+, A_-, \tilde{A}_5 \right)$,

$$\mathbb{P}\left[\sup_{s \in \mathcal{F}_{>C}} P_n (K s_M - K s) \geq (1 + A_5 \nu_n)^2 \frac{D_M - 1}{2n} \right] \leq 2n^{-\beta} \ . \tag{156}$$

Proof. The proof is similar to that of Lemma 23 and follows from the same kind of computations. First observe that

$$\begin{aligned}
&\sup_{s \in \mathcal{F}_{>C}} P_n (K s_M - K s) \\
&= \sup_{s \in \mathcal{F}_{>C}} \left\{ P_n \left(\psi_{1,M} \cdot (s_M - s) - \psi_2 \circ \left(\frac{s - s_M}{s_M} \right) \right) \right\} \\
&= \sup_{s \in \mathcal{F}_{>C}} \left\{ (P_n - P) \left(\psi_{1,M} \cdot (s_M - s) \right) - (P_n - P) \left(\psi_2 \circ \left(\frac{s - s_M}{s_M} \right) \right) - P (K s - K s_M) \right\} \\
&= \sup_{s \in \mathcal{F}_{>C}} \left\{ (P_n - P) \left(\psi_{1,M} \cdot (s_M - s) \right) - P (K s - K s_M) - (P_n - P) \left(\psi_2 \circ \left(\frac{s - s_M}{s_M} \right) \right) \right\} \\
&\leq \sup_{L > C} \sup_{s \in D_L} \left\{ (P_n - P) \left(\psi_{1,M} \cdot (s_M - s) \right) - (1 - L_{A_{\min}, A_\infty} \nu_n) L - (P_n - P) \left(\psi_2 \circ \left(\frac{s - s_M}{s_M} \right) \right) \right\} \text{ by (89)} \\
&\leq \sup_{L > C} \left\{ \sqrt{2L} \sqrt{\sum_{I \in \Lambda_M} (P_n - P)^2 (\varphi_I)} - (1 - L_{A_{\min}, A_\infty} \nu_n) L + \sup_{s \in \mathcal{F}_L} \left| (P_n - P) \left(\psi_2 \circ \left(\frac{s - s_M}{s_M} \right) \right) \right| \right\}
\end{aligned} \tag{157}$$

where the last bound follows from Cauchy-Schwarz inequality. From Lemma 18 and since for all $n \geq n_0 (A_-)$, $D_M \geq A_- (\ln n)^2 \geq 2$, we can deduce that for all $n \geq n_0 (A_-)$,

$$\mathbb{P}\left[\sqrt{\sum_{I \in \Lambda_M} (P_n - P)^2 (\varphi_I)} \geq (1 + L_{r_M, \beta} \nu_n) \sqrt{\frac{D_M - 1}{n}} \right] \leq n^{-\beta} \ . \tag{158}$$

Now, since

$$C \geq \frac{D_M}{2n}$$

we can apply Lemma 22 with $A_l = 1/2$, and deduce that, for all $n \geq n_0\left(A_{\min}, A_\infty, A_+\right)$,

$$\mathbb{P}\left[\forall L > C, \quad \sup_{s\in\mathcal{F}_L}\left|\left(P_n - P\right)\left(\psi_2 \circ \left(\frac{s - s_M}{s_M}\right)\right)\right| \geq L_{A_\infty, A_{\min}, A_-, \beta} \times \nu_n \sqrt{\frac{L\left(D_M - 1\right)}{n}}\right] \leq n^{-\beta} \tag{159}$$

Now using (158) and (159) in (157) we obtain, for all $n \geq n_0\left(A_{\min}, A_\infty, A_+, A_-\right)$,

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_{>C}} P_n\left(Ks_M - Ks\right) \geq \sup_{L>C}\left\{\left(1 + L_{A_\infty, r_M, A_{\min}, A_-, \beta} \times \nu_n\right)\sqrt{\frac{2L\left(D_M - 1\right)}{n}} - \left(1 - L_{r_M,\beta}\right)L\right\}\right] \leq 2n^{-\beta} \tag{160}$$

and we set $\tilde{A}_5 = L_{A_\infty, r_M, A_{\min}, A_-, \beta} \vee L_{r_M,\beta}$ where $L_{A_\infty, r_M, A_{\min}, A_-, \beta}$ and $L_{r_M,\beta}$ are the constants appearing in (160). Since, for all $n \geq n_0\left(A_+, A_-, \tilde{A}_5\right)$, $0 < \frac{1+\tilde{A}_5\nu_n}{1-\tilde{A}_5\nu_n} \leq 1 + 4\tilde{A}_5\nu_n$, and for $C \geq \left(1 + 4\tilde{A}_5\nu_n\right)^2 \frac{D_M - 1}{2n}$ we get by simple calculations, for all $n \geq n_0\left(A_+, A_-, \tilde{A}_5\right)$,

$$\sup_{L>C}\left\{\sqrt{2L}\left(1 + \tilde{A}_5\nu_n\right)\sqrt{\frac{D_M - 1}{n}} - \left(1 - \tilde{A}_5\nu_n\right)L\right\} = \left(1 + \tilde{A}_5\nu_n\right)\sqrt{\frac{2C\left(D_M - 1\right)}{n}} - \left(1 - \tilde{A}_5\nu_n\right)C \ .$$

Moreover, we have $C \leq A_u\frac{D_M}{n}$, so for all $n \geq n_0\left(A_-\right)$, $C \leq \sqrt{\frac{2A_uC(D_M-1)}{n}}$ and and as a consequence, for all $n \geq n_0\left(A_+, A_-, \tilde{A}_5\right)$,

$$\sup_{L>C}\left\{\sqrt{2L}\left(1 + \tilde{A}_5\nu_n\right)\sqrt{\frac{D_M - 1}{n}} - \left(1 - \tilde{A}_5\nu_n\right)L\right\} \leq \left(1 + \left(1 + \sqrt{A_u}\right)\tilde{A}_5\nu_n\right)\sqrt{\frac{2C\left(D_M - 1\right)}{n}} - C \ ,$$

so, for all $n \geq n_0\left(A_{\min}, A_\infty, A_+, A_-, \tilde{A}_5\right)$,

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_{>C}} P_n\left(Ks_M - Ks\right) \geq \left(1 + \left(1 + \sqrt{A_u}\right)\tilde{A}_5\nu_n\right)\sqrt{\frac{2C\left(D_M - 1\right)}{n}} - C\right] \leq 2n^{-\beta}$$

which gives the first part of the lemma by setting $A_5 = 4\tilde{A}_5 \vee \left(1 + \sqrt{A_u}\right)\tilde{A}_5$. The second part comes from (160) and the fact that, for any value of $C \geq 0$, for all $n \geq n_0\left(A_+, A_-, \tilde{A}_5\right)$,

$$\sup_{L>C}\left\{\sqrt{2L}\left(1 + \tilde{A}_5\nu_n\right)\sqrt{\frac{D_M - 1}{n}} - \left(1 - \tilde{A}_5\nu_n\right)L\right\} \leq \left(1 + 4\tilde{A}_5\nu_n\right)^2\frac{D_M - 1}{2n} \ .$$

■

Lemma 25 *Let $r > 1$ and $C, \beta > 0$. Assume that (**Abd**) and (**Alr**) hold. If positive constants A_-, A_+, A_l, A_u exist such that*

$$A_+\frac{n}{\left(\ln n\right)^2} \geq D_M \geq A_-\left(\ln n\right)^2 \quad \text{and} \quad A_l\frac{D_M}{n} \leq rC \leq A_u\frac{D_M}{n} \ ,$$

and if the constant A_∞ defined in (78) satisfies

$$A_\infty \geq 64B_2\sqrt{A_u}A_*r_M \ ,$$

then a positive constant $L_{A_-, A_l, A_u, A_{\min}, A_\infty, \beta}$ exists such that, for all $n \geq n_0\left(A_+, A_-, A_l, A_u, r_M, A_{\min}, A_\infty\right)$,

$$\mathbb{P}\left(\sup_{s\in\mathcal{F}_{(C,rC]}} P_n\left(Ks_M - Ks\right) \leq \left(1 - L_{A_-, A_l, A_u, A_{\min}, A_\infty, \beta} \times \nu_n\right)\sqrt{\frac{2rC\left(D_M - 1\right)}{n}} - rC\right) \leq 2n^{-\beta} \ ,$$

where $\nu_n = \max\left\{\sqrt{\frac{\ln n}{D_M}}, \ \sqrt{\frac{D_M \ln n}{n}}\right\}$.

Proof. Start with

$$\begin{aligned}
&\sup_{s\in\mathcal{F}_{(C,rC]}} P_n\left(Ks_M - Ks\right)\\
&= \sup_{s\in\mathcal{F}_{(C,rC]}} \left\{\left(P_n - P\right)\left(Ks_M - Ks\right) + P\left(Ks_M - Ks\right)\right\}\\
&\geq \sup_{s\in\mathcal{F}_{(C,rC]}} \left(P_n - P\right)\left(\psi_{1,M}\cdot\left(s_M - s\right)\right) - \sup_{s\in\mathcal{F}_{(C,rC]}} \left(P_n - P\right)\left(\psi_2\circ\left(\frac{s-s_M}{s_M}\right)\right) - \sup_{s\in\mathcal{F}_{(C,rC]}} P\left(Ks - Ks_M\right)\\
&\geq \sup_{s\in\mathcal{F}_{(C,rC]}} \left(P_n - P\right)\left(\psi_{1,M}\cdot\left(s_M - s\right)\right) - \sup_{s\in\mathcal{F}_{rC}} \left(P_n - P\right)\left(\psi_2\circ\left(\frac{s-s_M}{s_M}\right)\right) - rC
\end{aligned} \tag{161}$$

and set

$$\begin{aligned}
S_{1,r,C} &= \sup_{s\in\mathcal{F}_{(C,rC]}} \left(P_n - P\right)\left(\psi_{1,M}\cdot\left(s_M - s\right)\right)\\
M_{1,r,C} &= \mathbb{E}\left[\sup_{s\in\mathcal{F}_{(C,rC]}} \left(P_n - P\right)\left(\psi_{1,M}\cdot\left(s_M - s\right)\right)\right]\\
b_{1,r,C} &= \sup_{s\in\mathcal{F}_{(C,rC]}} \left\|\psi_{1,M}\cdot\left(s_M - s\right) - P\psi_{1,M}\cdot\left(s_M - s\right)\right\|_\infty\\
\sigma^2_{1,r,C} &= \sup_{s\in\mathcal{F}_{(C,rC]}} \mathrm{Var}\left(\psi_{1,M}\cdot\left(s_M - s\right)\right)\ .
\end{aligned}$$

By Klein-Rio's Inequality (175), we get, for all $\delta, x > 0$,

$$\mathbb{P}\left(S_{1,r,C} \leq \left(1-\delta\right)M_{1,r,C} - \sqrt{\frac{2\sigma^2_{1,r,C}x}{n}} - \left(1+\frac{1}{\delta}\right)\frac{b_{1,r,C}x}{n}\right) \leq \exp\left(-x\right). \tag{162}$$

Then, notice that all conditions of Lemma 19 are satisfied and that it gives for all $n \geq n_0\left(A_+, A_-, A_l, A_u, r_M, A_{\min}\right)$,

$$\mathbb{E}\left[\sup_{s\in\mathcal{F}_{(C,rC]}} \left(P_n - P\right)\left(\psi_{1,M}\cdot\left(s_M - s\right)\right)\right] \geq \left(1 - \frac{L_{A_l,A_u,A_{\min}}}{\sqrt{D_M}}\right)\sqrt{\frac{2rC\left(D_M - 1\right)}{n}}\ . \tag{163}$$

In addition, observe that

$$\sigma^2_{1,r,C} \leq \sup_{s\in\mathcal{F}_{(C,rC]}} P\left(\psi^2_{1,M}\left(s_M - s\right)^2\right) \leq 2rC \tag{164}$$

and for all $n \geq n_0\left(A_+, A_{\min}, A_\infty\right)$,

$$b_{1,r,C} \leq 2\sup_{s\in\mathcal{F}_{(C,rC]}} \left\|\psi_{1,M}\cdot\left(s_M - s\right)\right\|_\infty \leq 2A^{-1}_{\min}\tilde{R}_{n,D_M,\alpha} \leq 1 \tag{165}$$

Hence, using (163), (164) and (165) in Inequality (162), we get for all $x > 0$ and all $n \geq n_0\left(A_+, A_-, A_l, A_u, r_M, A_{\min}, A_\infty\right)$,

$$\mathbb{P}\left(S_{1,r,C} \leq \left(1-\delta\right)\left(1 - \frac{L_{A_l,A_u,A_{\min}}}{\sqrt{D_M}}\right)\sqrt{\frac{2rC\left(D_M-1\right)}{n}} - \sqrt{\frac{4rCx}{n}} - \left(1+\frac{1}{\delta}\right)\frac{x}{n}\right) \leq \exp\left(-x\right)\ .$$

Now, taking $x = \beta\ln n$, $\delta = \sqrt{\frac{\ln n}{D_M}}$, we can deduce by simple computations that a positive constant $L_{A_l,A_u,A_{\min},\beta}$ exists such that, for all $n \geq n_0\left(A_+, A_-, A_l, A_u, r_M, A_{\min}, A_\infty\right)$,

$$\mathbb{P}\left(S_{1,r,C} \leq \left(1 - L_{A_l,A_u,A_{\min},\beta}\sqrt{\frac{\ln n}{D_M}}\right)\sqrt{\frac{2rC\left(D_M-1\right)}{n}}\right) \leq n^{-\beta} \tag{166}$$

and as

$$\sqrt{\frac{\ln n}{D_M}} \leq \nu_n \ ,$$

(166) gives, for all $n \geq n_0 \left(A_+, A_-, A_l, A_u, r_M, A_{\min}, A_\infty\right)$,

$$\mathbb{P}\left(S_{1,r,C} \leq \left(1 - L_{A_l, A_u, A_{\min}, \beta} \nu_n\right) \sqrt{\frac{2rC\left(D_M - 1\right)}{n}}\right) \leq n^{-\beta} \ . \tag{167}$$

Moreover, from Lemma 22 we can deduce that, for all $n \geq n_0 \left(A_{\min}, A_\infty, A_+, A_l\right)$,

$$\mathbb{P}\left[\sup_{s \in \mathcal{F}_{rC}} \left|\left(P_n - P\right)\left(\psi_2 \circ \left(\frac{s - s_M}{s_M}\right)\right)\right| \geq L_{A_-, A_l, A_{\min}, \beta} \sqrt{\frac{rC\left(D_M - 1\right)}{n}} \tilde{R}_{n, D_M, \alpha}\right] \leq n^{-\beta} \tag{168}$$

and noticing that

$$\tilde{R}_{n,D,\alpha} = A_\infty \sqrt{\frac{D \ln n}{n}} \leq A_\infty \nu_n$$

we deduce from (168) that, for all $n \geq n_0 \left(A_{\min}, A_\infty, A_+, A_l\right)$,

$$\mathbb{P}\left[\sup_{s \in \mathcal{F}_{rC}} \left|\left(P_n - P\right)\left(\psi_2 \circ \left(\frac{s - s_M}{s_M}\right)\right)\right| \geq L_{A_-, A_l, A_{\min}, A_\infty, \beta} \nu_n \sqrt{\frac{2rC\left(D_M - 1\right)}{n}}\right] \leq n^{-\beta} \ . \tag{169}$$

Finally, using (167) and (169) in (161) we get that, for all $n \geq n_0 \left(A_+, A_-, A_l, A_u, r_M, A_{\min}, A_\infty\right)$,

$$\mathbb{P}\left(\sup_{s \in \mathcal{F}_{(C, rC]}} P_n\left(Ks_M - Ks\right) \leq \left(1 - L_{A_-, A_l, A_u, A_{\min}, A_\infty, \beta} \times \nu_n\right) \sqrt{\frac{2rC\left(D_M - 1\right)}{n}} - rC\right) \leq 2n^{-\beta},$$

which concludes the proof. ■

5.6 Probabilistic Tools

We recall here the main probabilistic results that are instrumental in our proofs.
Let us begin with the L_p-version of Hoffmann-Jørgensen's inequality, that can be found for example in [21], Proposition 6.10, p.157.

Theorem 26 *For any independent mean zero random variables Y_j, $j = 1, ..., n$ taking values in a Banach space $(\mathcal{B}, \|.\|)$ and satisfying $\mathbb{E}\left[\|Y_j\|^p\right] < +\infty$ for some $p \geq 1$, we have*

$$\mathbb{E}^{1/p}\left\|\sum_{j=1}^n Y_j\right\|^p \leq B_p\left(\mathbb{E}\left\|\sum_{j=1}^n Y_j\right\| + \mathbb{E}^{1/p}\left(\max_{1 \leq j \leq n} \|Y_j\|\right)^p\right)$$

where B_p is a universal constant depending only on p.

We will use this theorem for $p = 2$ in order to control suprema of empirical processes. In order to be more specific, let $\mathcal{F}$ be a class of measurable functions from a measurable space $\mathcal{Z}$ to $\mathbb{R}$ and $(X_1, ..., X_n)$ be independent variables of common law P taking values in $\mathcal{Z}$. We then denote by $\mathcal{B} = l^\infty\left(\mathcal{F}\right)$ the space of uniformly bounded functions on $\mathcal{F}$ and, for any $b \in \mathcal{B}$, we set $\|b\| = \sup_{f \in \mathcal{F}} |b\left(f\right)|$. Thus $(\mathcal{B}, \|.\|)$ is a Banach space. Indeed we shall apply Theorem 26 to the independent random variables, with mean zero and taking values in $\mathcal{B}$, defined by

$$Y_j = \left\{f\left(X_j\right) - Pf, \ f \in \mathcal{F}\right\} \ .$$

More precisely, we will use the following result, which is a straightforward application of Theorem 26. Denote by

$$P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$$

the empirical measure associated to the sample $(X_1, ..., X_n)$ and by

$$\|P_n - P\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |(P_n - P)(f)|$$

the supremum of the empirical process over $\mathcal{F}$.

Corollary 27 *If $\mathcal{F}$ is a class of measurable functions from a measurable space $\mathcal{Z}$ to $\mathbb{R}$ satisfying*

$$\sup_{z \in \mathcal{Z}} \sup_{f \in \mathcal{F}} |f(z) - Pf| = \sup_{f \in \mathcal{F}} \|f - Pf\|_{\infty} < +\infty$$

and $(X_1, ..., X_n)$ are n i.i.d. random variables taking values in $\mathcal{Z}$, then an absolute constant B_2 exists such that,

$$\mathbb{E}^{1/2}\left[\|P_n - P\|_{\mathcal{F}}^2\right] \leq B_2 \left(\mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right] + \frac{\sup_{f \in \mathcal{F}} \|f - Pf\|_{\infty}}{n}\right). \tag{170}$$

Another tool we need is a comparison theorem for Rademacher processes, see Theorem 4.12 of [21]. A function $\varphi : \mathbb{R} \to \mathbb{R}$ is called a contraction if $|\varphi(u) - \varphi(v)| \leq |u - v|$ for all $u, v \in \mathbb{R}$. Moreover, for a subset $T \subset \mathbb{R}^n$ we set

$$\|h(t)\|_T = \|h\|_T = \sup_{t \in T} |h(t)| .$$

Theorem 28 *Let $(\varepsilon_1, ..., \varepsilon_n)$ be n i.i.d. Rademacher variables and $F : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ be a convex and increasing function. Furthermore, let $\varphi_i : \mathbb{R} \longrightarrow \mathbb{R}$, $i \leq n$, be contractions such that $\varphi_i(0) = 0$. Then, for any bounded subset $T \subset \mathbb{R}^n$,*

$$\mathbb{E}F\left(\left\|\sum_i \varepsilon_i \varphi_i(t_i)\right\|_T\right) \leq 2\mathbb{E}F\left(\left\|\sum_i \varepsilon_i t_i\right\|_T\right).$$

The next tool is the well known Bernstein's inequality, that can be found for example in [23], Proposition 2.9.

Theorem 29 *(Bernstein's inequality) Let $(X_1, ..., X_n)$ be independent real valued random variables and define*

$$S = \frac{1}{n}\sum_{i=1}^{n} (X_i - \mathbb{E}[X_i]).$$

Assuming that

$$v = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[X_i^2\right] < \infty$$

and

$$X_i \leq b \quad a.s.$$

we have, for every $x > 0$,

$$\mathbb{P}\left[|S| \geq \sqrt{2v\frac{x}{n}} + \frac{bx}{3n}\right] \leq 2\exp(-x). \tag{171}$$

We now turn to concentration inequalities for the empirical process around its mean. Bousquet's inequality [12] provides optimal constants for the deviations above the mean. Klein-Rio's inequality [18] gives sharp constants for the deviations below the mean, that slightly improves Klein's inequality [19].

Theorem 30 *Let $(\xi_1, ..., \xi_n)$ be n i.i.d. random variables having common law P and taking values in a measurable space $\mathcal{Z}$. If $\mathcal{F}$ is a class of measurable functions from $\mathcal{Z}$ to $\mathbb{R}$ satisfying*

$$|f(\xi_i) - Pf| \leq b \quad a.s., \text{ for all } f \in \mathcal{F},\ i \leq n,$$

then, by setting

$$\sigma_{\mathcal{F}}^2 = \sup_{f \in \mathcal{F}} \left\{P\left(f^2\right) - (Pf)^2\right\},$$

we have, for all $x \geq 0$,

Bousquet's inequality *:*

$$\mathbb{P}\left[\|P_n - P\|_{\mathcal{F}} - \mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right] \geq \sqrt{2\left(\sigma_{\mathcal{F}}^2 + 2b\mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right]\right)\frac{x}{n}} + \frac{bx}{3n}\right] \leq \exp\left(-x\right) \tag{172}$$

and we can deduce that, for all $\varepsilon, x > 0$*, it holds*

$$\mathbb{P}\left[\|P_n - P\|_{\mathcal{F}} - \mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right] \geq \sqrt{2\sigma_{\mathcal{F}}^2\frac{x}{n}} + \varepsilon\mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right] + \left(\frac{1}{\varepsilon} + \frac{1}{3}\right)\frac{bx}{n}\right] \leq \exp\left(-x\right). \tag{173}$$

Klein-Rio's inequality :

$$\mathbb{P}\left[\mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right] - \|P_n - P\|_{\mathcal{F}} \geq \sqrt{2\left(\sigma_{\mathcal{F}}^2 + 2b\mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right]\right)\frac{x}{n}} + \frac{bx}{n}\right] \leq \exp\left(-x\right) \tag{174}$$

and again, we can deduce that, for all $\varepsilon, x > 0$*, it holds*

$$\mathbb{P}\left[\mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right] - \|P_n - P\|_{\mathcal{F}} \geq \sqrt{2\sigma_{\mathcal{F}}^2\frac{x}{n}} + \varepsilon\mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right] + \left(\frac{1}{\varepsilon} + 1\right)\frac{bx}{n}\right] \leq \exp\left(-x\right). \tag{175}$$

The following result is due to Ledoux [20]. We will use it along the proofs through Corollary 32 which is sated below. From now on, we set for short $Z = \|P_n - P\|_{\mathcal{F}}$.

Theorem 31 *Let* $(\xi_1, ..., \xi_n)$ *be independent random with values in some measurable space* $(\mathcal{Z}, \mathcal{T})$ *and* $\mathcal{F}$ *be some countable class of real-valued measurable functions from* $\mathcal{Z}$*. Let* $\left(\xi_1', ..., \xi_n'\right)$ *be independent from* $(\xi_1, ..., \xi_n)$ *and with the same distribution. Setting*

$$v = \mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\left(f\left(\xi_i\right) - f\left(\xi_i'\right)\right)^2\right]$$

then

$$\mathbb{E}\left[Z^2\right] - \mathbb{E}\left[Z\right]^2 \leq \frac{v}{n}\ .$$

Corollary 32 *Under notations of Theorem 30, if some* $\varkappa_n \in (0,1)$ *exists such that*

$$\varkappa_n^2\mathbb{E}\left[Z^2\right] \geq \frac{\sigma^2}{n}$$

and

$$\varkappa_n^2\sqrt{\mathbb{E}\left[Z^2\right]} \geq \frac{b}{n}$$

then we have, for a numerical constant $A_{1,-}$*,*

$$\left(1 - \varkappa_n A_{1,-}\right)\sqrt{\mathbb{E}\left[Z^2\right]} \leq \mathbb{E}\left[Z\right]\ .$$

Proof of Corollary 32. Just use Theorem 31, noticing the fact that

$$\sqrt{\mathbb{E}\left[Z^2\right]} - \mathbb{E}\left[Z\right] \leq \sqrt{\mathbb{V}\left(Z\right)}$$

and that, with notations of Theorem 31,

$$v \leq 2\sigma^2 + 32b\mathbb{E}\left[Z\right]\ .$$

The result then follows from straightforward calculations. ■

References

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

[2] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Trans. Automatic Control*, AC-19:716–723, 1974. System identification and time-series analysis.

[3] Sylvain Arlot. *Resampling and Model Selection*. PhD thesis, University Paris-Sud 11, December 2007. oai:tel.archives-ouvertes.fr:tel-00198803_v1.

[4] Sylvain Arlot. Model selection by resampling penalization, March 2008. oai:hal.archives-ouvertes.fr:hal-00262478_v2.

[5] Sylvain Arlot. V-fold cross-validation improved: V-fold penalization, February 2008. arXiv:0802.0566v2.

[6] Sylvain Arlot and Pascal Massart. Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, 10:245–279 (electronic), 2009.

[7] A. R. Barron. Limits of information, markov chains, and projections. In *Proceedings. 2000 International Synopsium on Information Theory*, page 25, 2000.

[8] Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.

[9] A.R. Barron and C.H. Sheu. Approximation of density functions by sequences of exponential families. *Ann. Statist.*, 19(3):1347–1369, 1991.

[10] Lucien Birgé and Pascal Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007.

[11] S. Boucheron and P. Massart. A high dimensional Wilks phenomenon. *Probability Theory and Related Fields*, 2010. To appear.

[12] Olivier Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, 334(6):495–500, 2002.

[13] P. Burman. Estimation of equifrequency histogram. *Statist. Probab. Lett.*, 56(3):227–238, 2002.

[14] G. Castellan. Modified Akaike's criterion for histogram density estimation. *Technical report ♯99.61, Université de Paris-Sud.*, 1999.

[15] Gwénaëlle Castellan. Density estimation via exponential model selection. *IEEE Trans. Inform. Theory*, 49(8):2052–2060, 2003.

[16] Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.*, 3(1):146–158, 1975.

[17] Imre Csiszár and František Matúš. Information projections revisited. *IEEE Trans. Inform. Theory*, 49(6):1474–1490, 2003.

[18] R. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *Annals of Probability*, 1:63–87 (electronic), 2005.

[19] T. Klein. Une inégalité de concentration à gauche pour les processus empiriques. *C.R. Acad. Sci. Paris, Ser I*, 334:500–505, 2002.

[20] M. Ledoux. On Talagrand's deviation inequalities for product measures. *ESAIM: Probability and Stattistics*, 1:63–87, 1996.

[21] M. Ledoux and M. Talagrand. *Probability in Banach spaces*. Springer, Berlin, 1991.

[22] Colin L. Mallows. Some comments on C_p. *Technometrics*, 15:661–675, 1973.

[23] P. Massart. *Concentration Inequalities and Model Selection*. Springer-Verlag, 2007.

[24] Adrien Saumard. Optimal upper and lower bounds for the true and empirical excess risks in heteroscedastic least-squares regression, August 2010. hal-00512304, v1.

[25] Adrien Saumard. The slope heuristics in heteroscedastic regression, August 2010. hal-00512306, v1.

[26] Charles J. Stone. An asymptotically optimal histogram selection rule. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983)*, Wadsworth Statist./Probab. Ser., pages 513–520, Belmont, CA, 1985. Wadsworth.

[27] Charles J. Stone. Uniform error bounds involving logspline models. In *Probability, statistics, and mathematics*, pages 335–355. Academic Press, Boston, MA, 1989.

[28] S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.*, 9(3):60–62, 1938.